EyeGaze

Facilitating eye contact in a video mediated setting

By: Anne Kathrine Jensen
Thomas Søndersø Nielsen
Jacob Haubach Smedegård

Résumé

In this Master Thesis we present a novel approach to video mediated communication called EyeGaze. EyeGaze facilitates eye contact in a conversation by rendering a person from a virtual perspective based on head tracking data of the observer. In a conventional video communication system, the camera is forced to be placed on top of, or to the side of the screen, due to the opaque nature of both the camera and the screen technologies. In order to obtain a sense of mutual eye contact, you would be required to look at two places at once, a feat most people would find challenging. EyeGaze solves this problem by using 3D depth sensing cameras placed strategically around the screen in order to capture the person in full real-time 3D. The 3D model is generated by merging depth frames from each camera into a single 3D model. We represent the 3D model as a voxel array in GPU memory, which facilitates merging new frames easily, as well as providing tolerance against missing data in the depth frame.

Our main contribution spans two articles, one detailing the technical aspects of the EyeGaze implementation called "*EyeGaze: Eye Contact and Gaze Awareness in Video*" and one detailing a within-subject study of the EyeGaze system called "*EyeGaze: A Comparative Analysis of Video Mediated Eye Contact*". Our technical article presents the EyeGaze system and frames it in related literature on video mediated communication and advancements made to facilitate nonverbal cues in such settings. The article then details the EyeGaze system and the implementational aspects of our prototype. Finally a formative assessment of the performance and quality of EyeGaze is presented.

The second article presents a comparative study done on the EyeGaze system. We summarise the literature pertaining to the role of eye gaze in communication and studies which explore the effects of eye contact in virtual avatar representations. We then describe our within-subject study, the procedure, and the questionnaire which the participants answered directly after each session. Our findings are presented, showing that the EyeGaze prototype results in a significantly better sense of eye contact between the participants compared to a Skype video conversation. Other findings points to a better sense of involvement and turn-taking when using EyeGaze and a general positive bias towards EyeGaze in the responses.

Introduction

Human communication is more than simply speech. When people communicate face to face, eye contact, gaze direction, body posture, gestures, and relative distances between persons are all indicators used actively in the communication. However, when communicating using technology such as the telephone, or video links (i.e. Skype or similar IP solutions), many of these supporting actions are lost. Understanding and designing the use of video links for distributed teams has been the focus of research for close to 30 years. Efforts have been made to reintroduce the experience of gaze direction, eye contact, and gestures in these systems with the main focus being on eye contact. Solutions to the problem have ranged from the simple mirror box to creating rooms designed specifically to give the appearance of a blended space.

In this thesis, we introduce a new approach building upon previous concepts. We virtually recreate reality using depth sensing cameras to capture the environment. Our approach differs from other contemporary systems such as the Encumberance Free Telepresence System presented in [3] by our use of a persistent voxel representation which allows us to handle input noise without the need for preprocessing of the input. We make use of advances made in the field of 3D reconstruction, using a technique for recreating scanned 3D objects from an inaccurate and noisy source, such as a depth sensing camera. This technique forms the basis for our proof-of-concept prototype, EyeGaze, which recreates the environment in 3D, and allows for an arbitrary viewing position.

This Master Thesis is the resulting work of the implementation and study of EyeGaze. We begin by describing the major research contributions in this thesis, followed by a recount and reflection of the first study we performed to evaluate EyeGaze. We then move on to the two papers, which form the main part of this thesis. Paper one describes the implementation of EyeGaze, as well as a formative evaluation of the quality and performance. In the second paper, we evaluate the user experience of EyeGaze through an empirical study, and compare it with conventional face-to-face communication, as well as a contemporary video chat solution, Skype.

Research Contributions

The two papers comprising the majority of this thesis have been written based on research done during our 10th semester. Both papers are written to be stand-alone, however, the second paper makes use of our prototype described in the first paper. A chronological read-through is therefore recommended.

The articles are prefaced with a chapter describing a between-subject study that we performed on EyeGaze. The findings of this study points to a serious problem when using a between-subject approach to understand the relative qualitative merits of these conditions. In none of the questionnaire measures did a face-to-face conversation outperform the basic video link or EyeGaze significantly. Our discussion of the findings point to problems in the applied method. The study found in the second article of this thesis was designed with these issues in mind and did achieve significant difference between conditions in some measures.

The first paper presents and discusses the technical implementation of our proof-of-concept prototype, EyeGaze, which has been largely rewritten and optimised. In addition, EyeGaze has been extended to support multiple cameras, head tracking, network communication, texturing from multiple cameras, and a high resolution voxel representations since the prototype developed during our 9th semester. We present an overview of literature on the video-mediated communication approach from its start at Xerox PARC nearly 30 years ago up to advancements in blended interaction spaces, such as BISi by Paay et al. [4], and the advanced 3D solutions of Maimone and Fuchs [3]. The necessary implementational details for recreating a similar system are described, as well as the reasoning for our choices in both technology and method. Our main research contribution in this paper is the novel use of a volumetric representation to merge raw depth input from multiple Kinect sensors into a single 3D model of a mobile individual and how this can be used in a video mediated conversation. This approach originates from advances in 3D scanning using commodity hardware, such as the Kinect sensor, and storing the merged volumetric representation persistently on the GPU as a voxel grid. We detail how it is possible to utilise the parallelised nature of the GPU to achieve acceptable framerates when merging and rendering the scene, even on average commodity graphics cards. To help facilitate the experience of eye contact, EyeGaze renders video from a perspective obtained by tracking the head of the user. We have experimented with rendering using textures from a single Kinect, as well as textures from multiple Kinects. Lastly, we show rendered images confirming that our system is indeed able to facilitate the experience of eye contact and give some performance numbers showing interactive update rates for the video.

The second paper describes a within-subject study of the EyeGaze system, which describes a comparative analysis of EyeGaze with contemporary video mediated communication. This work is framed in literature on eye contact and face gaze, as well as other empirical studies performed to evaluate the importance of eye contact. The study had 30 participants and performed in a within-subject manner with three conditions differing in the use of video technology to mediate the conversation. The

three conditions were a normal face to face condition without technology involved, a Skype condition, consisting of a single webcam placed above a screen for each participant, and the EyeGaze condition consisting of an EyeGaze setup for each participant. Our main contribution in this paper is our findings from the study, which show consistently better ratings for EyeGaze than Skype, though not all were statistically significant. When asked directly about their experience of achieving eye contact with their conversation partner, participants overwhelmingly preferred EyeGaze compared to Skype. Another measure which showed a strong preference between the two was Involvement. Furthermore, some measures showed a statistical significance between Face-to-Face and Skype, but not Face-to-Face and EyeGaze.

Reflection on Study Method

As part of our 10th semester we performed a between-subject study of EyeGaze with the intention of comparing it against a contemporary video mediated communication methods and a normal face to face conversation. The users evaluated their experience communicating in one of three conditions (Face-to-Face, EyeGaze, and a Basic Video Link) by answering a questionnaire in which a series of statements were to be evaluated on a nine-point Likert-scale. The findings of the study were surprisingly inconclusive in respect to the conditions. We barely found any subject in which there was a significant difference in the means of users' answers between conditions. Neither analysing the results of the questionnaire in a Likert-type method nor as a Likert-scale gave any significant results. This could either indicate that there is no remarkable difference between communicating face-to-face, via a basic video link, or through EyeGaze, or that the methodology used influenced the results. As we consider it highly unlikely that the three conditions give the same experience, we decided to re-evaluate our approach on the design of the study. In the following we present the design and findings of the between-subject study we performed, focusing on reflection of how and why the study showed such unsatisfactory results, and conclude with what we learned. These observations and thoughts formed the basis of our within-subject study presented in "*EyeGaze: A Comparative Analysis of Video Mediated Eye Contact*".

Study Design

Our initial goal was to isolate the conditions and ensure that the users' experiences were not unduly influenced between the conditions, which lead us to choose the between-subject method inspired by Garau et al. [1] over the within-subject method. The work by Garau et al. showed a significant difference in parts of users' experience when discussing through a video-audio link compared with when interacting a virtual avatar whose gazing behaviour and head movements were either random or based on real-time eye tracking data. In their study, Garau et al. performed a between-subject study with 100 participants, each of whom answered a questionnaire after the study.

We designed our study similar to that of Garau et al. in scale and setup. For our between-subject study we recruited 90 participants, who were randomly paired a partner and condition.

Population

Participants were recruited by sending e-mails to the students in the School of Information and Communication Technology, Department of Mechanical and Manufacturing Engineering, Department of

Mathematical Sciences, and Department of Communication and Psychology. In addition, we used Facebook, door-to-door recruitment, and word of mouth to ask for participants. All participants except one were in their 20's and all spoke Danish fluently. In an attempt to avoid pairs of participants who knew each other, we did not pair participants from the same year together, nor participants who had volunteered at approximately the same time. When doing more ad-hoc recruitment similar efforts were made to ensure that paticipants did not know each other.

Independent Variable

The three conditions in our between-subject study were chosen to allow us to compare the experience of communicating face-to-face and through contemporary video communications method, with our system, EyeGaze. To represent the contemporary video communication method, we chose to use a basic video link from the Kinect camera placed above the screen. The conditions are described in the following.

EyeGaze
: Each user sat at an EyeGaze setup, which consisted of three Microsoft Kinects and a 42" widescreen LCD screen. The setups were back to back, but were separated by a temporary wall.

Face-to-Face
: The participants were seated on opposite sides of a table. Approximately one meter above the center of the table, such that it did not obstruct the line of sight for the two participants, two cameras were placed pointing in opposite directions, each recording a single participant.

Basic Video Link
: Two 42" widescreen LCD displays were placed on opposite sides of a temporary wall, each with a single camera placed on top of it, pointing down in a 16°angle. The raw video was streamed in 30 FPS to the screen on the opposite side.

The EyeGaze condition and the Basic Video Link condition was performed in the same room. The EyeGaze condition was typically run before noon, after which the setup was switched to the Basic Video Link in the afternoon. The Face-to-Face conditon was run in a different room in parallel with the other two conditions.

Procedure

For each session, the test leader read aloud an introductory text explaining the session and the goals for the two participants. The participants were asked to sign a consent form, and were seated at the condition setup. A video recording of each participant was started and the participants were left to the assignment. The two participants then read through the four different conversational topics, and negotiated which topic they would discuss and who would argue in favour and who in opposition. The topics available were as follows

- Nuclear power in Denmark
- Assisted suicide
- A ban on fighting dogs
- Mandatory usage of bicycle helmets

Once the test subjects had reached an agreement, they discussed the subject for 8-10 minutes after which the test leader halted the discussion and changed the subject to that of death penalty in the Danish legal system. This discussion then carried on for the remaining 10 minutes. In cases where the parties were unable to discuss the chosen subject for the alloted 10 minutes, a new subject was negotiated and discussed. If all subjects were discussed and less than 20 minutes had passed, the session was halted early. The users then answered the following questionnaire. For each item in

the questionnaire, users were asked to evaluate the condition they were assigned to on a nine-point Likert-scale from completely disagree to completely agree.

Communication

1. I could easily tell when my partner was listening to me
2. I was able to take control of the conversation when I wanted to
3. It was easy for me to contribute to the conversation
4. The conversation seemed highly interactive
5. There were frequent and inappropriate interruptions
6. This felt like a natural conversation

Turn-Taking

7. I feel I interrupted my conversation partner often
8. I was often unsure of when my conversation partner was done talking

Involvement

9. I found it easy to keep track of the conversation
10. I felt completely absorbed in the conversation
11. I was easily distracted from the conversation

Co-Presence

12. I had a real sense of personal contact with my conversation partner
13. I was very aware of my conversation partner

Partner Evaluation

14. My partner was friendly
15. My partner did not take a personal interest in me
16. I trusted my partner
17. I enjoyed talking to my partner

Attentiveness

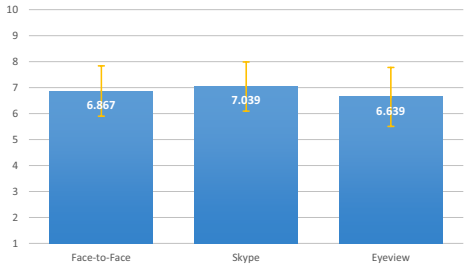
18. My conversation partner seemed attentive
19. My conversation partner listened to me

Findings and Discussion

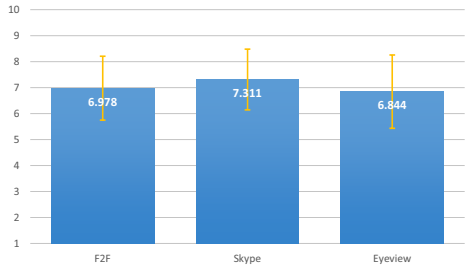
Our findings from the questionnaires are summarised in Figure 1. It shows a very surprising and not very informative tendency to rate each condition very similarly. Only the Involvement measure showed significant difference, $F(2, 87) = 3.76, p < 0.05$. Here the Basic Video Feed was rated significantly better than Face-to-Face ($p < 0.05$) , which seems to be at odds with any reasonable expectation. This is a highly unlikely result, and points to an issue in the methodology used. Our findings show that the participants gave their conditions a mean score of around seven on the nine-point scale. This highly consistent rating points to different scales being applied to the conditions, where each condition is rated against the expectations of that specific condition. These unusual results persist even when we look at each question invididually. In most cases we do not see any deviation from the general tendency to rate each condition the same. An illustrative example of this problem is the rating of candy. Picture a study is performed on the taste of candy. One participant is given one type of candy, and another participant is given a different type of candy. When asked, they both rate the taste as "okay", but this does not inform us of which candy is better.

This is the shortcoming of using a between-subject study for this line of questioning. Ultimately trying to eliminate any kind of comparison between the different conditions results in each participant rating things according to a personal scale. It seems likely that people have rated each conditions on its own merits. A Face-to-Face condition would clearly outperform the Basic Video Link and the EyeGaze condition, if you were to compare them, but on its own, it is simply a comparison against itself and what immediate improvements the person can think of. Usoh et al. [5] also note that when using virtual environment questionnaires in real life, even obvious questions can receive a lower score than what would be expected. A question such as whether the person has a sense of “being there” can result in a score lower than the maximum simply because people reinterpret the meaning of the question depending on the context in which it is asked.

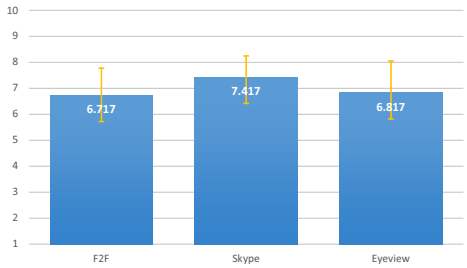
The problem with relating conditions to themselves is also present in regards to the basic video link. People are used to Skype which is quite similar to the basic video link. This can cause people to rate



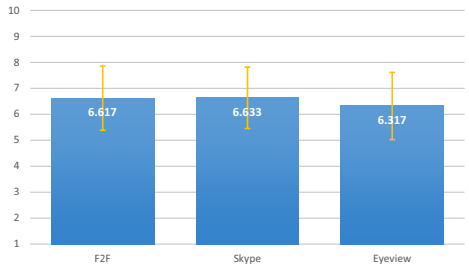
(a) Means for Communication measure



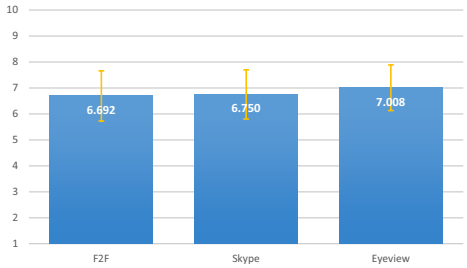
(b) Means for Turn-Taking measure



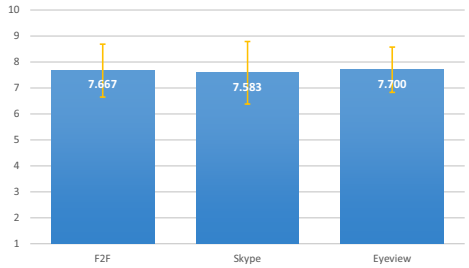
(c) Means for Involvement measure



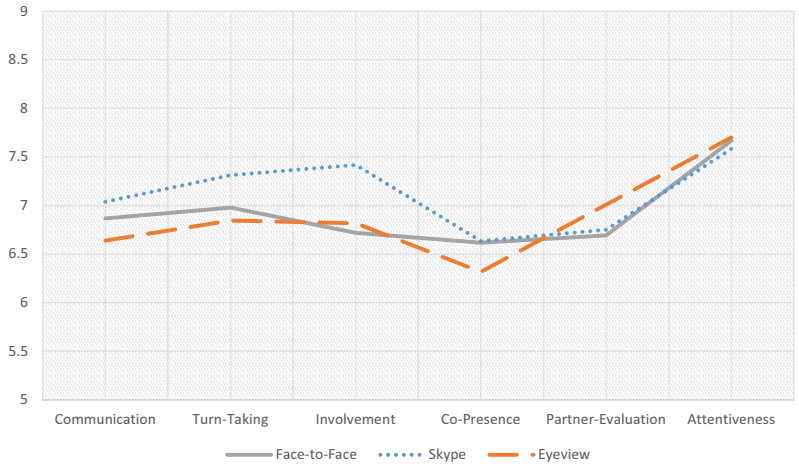
(d) Means for Co-Presence measure



(e) Means for Partner Evaluation measure



(f) Means for Attentiveness measure



(g) Means for all measures

Figure 1: Mean values for each measure for each condition, showing a very consistent rating between all measures and conditions

the Basic Video Link in regards to Skype, and thus the experience is actually good. With the EyeGaze condition people have no direct point of comparison, since they most likely have never interacted with such a system beforehand. The performance could in their mind probably be better, but it could certainly also be worse. Being uncertain of how to rate the system is certainly not a reason to rate it

badly, and being unfamiliar with the system, would most likely push the rating higher as a general courtesy.

Both users were recorded in all sessions, giving us approximately 30 hours of video. All the videos were manually analysed for eye movements, speech patterns such as turn-taking, the use of supporting statements such as utterances of "uhuh", "yes", "mhm", and any breakdown in communication between the two participants. The videos were evenly distributed between the three authors with each person receiving 10 hours of video distributed evenly between the three conditions. The analysis of the videos took approximately 135 hours. Looking at the detailed analysis of the video analysis data showed that gazing patterns and speech patterns were highly individual with mutual gaze in the sessions lasting between 2% - 66% of the total session length. Similar diversity was found in the number of support statements, which ranged between 1 and 72 in a single session. When looking at our data and our distribution of test subjects, we fear that individual differences in personality and age played a part in the uneven spread of the data, due to the between-subject design.

Based on these observations we designed a within-subject study presented in "*EyeGaze: A Comparative Analysis of Video Mediated Eye Contact*", focussing on the direct comparison between conditions and forces the participants to make an active choice of comparing each condition to the next.

Conclusion

This Master Thesis presented, framed, and studied a new approach to video mediated commmunication based on the 3D scanning approach of Izadi et al. [2]. Our novel system, EyeGaze, presented in "*EyeGaze: Eye Contact and Gaze Awareness in Video*", captures the user's environment in 3D in real-time, allowing a virtual camera to be positioned wherever provides the best experience of eye contact. Furthermore, EyeGaze tracks the user's head to continually provide the correct virtual camera placement. This also allows EyeGaze to give the effect of looking around corners, through the use of an asymmetric virtual camera.

The system was framed in literature related to video links, from the early systems studied at Xerox PARC to more advanced systems such as BISi [4] and the Encumberance Free Telepresence System [3]. We describe in detail how, through the use of a volumetric representation, our system can merge the output from multiple Kinects into a single model. This captures more details of the environment than a single Kinect, and allows for a robustness against sudden gaps or erroneous data. Furthermore, we provide a formative assessment of the model and render quality, and the performance of the implementation. We have provided a discussion of the results and pointed towards future improvements in the system.

To evaluate our system, we performed a comparative within-subject study, in "*EyeGaze: A Comparative Analysis of Video Mediated Eye Contact*", of the EyeGaze system against a natural face-to-face conversation and Skype. 30 participants were recruited and randomly paired. We present the results from our study which was based on questionnaire responses from all 30 participants, where they were asked to rate each question in relation to each condition, and judge whether they found Skype or EyeGaze more pleasing in relation to the question.

Our study showed three trends: *1*) EyeGaze was rated significantly better than Skype; *2*) EyeGaze was not rated significantly worse than Face-to-Face, while Skype was rated significantly worse than Face-to-Face, and finally; *3*) both Skype and EyeGaze were rated significantly worse than Face-to-Face.

Our findings show that users rated EyeGaze significantly higher than Skype when asked if they felt they had a good sense of eye contact. This is an important result, as facilitating an experience of eye contact is the primary goal in the design of EyeGaze. In addition our study showed a general tendency towards a better experience in EyeGaze compared to Skype, however, this tendency was not always significant. Both EyeGaze and Skype performed significantly worse than Face-to-Face, which was expected.

Research Paper 1

EyeGaze: Eye Contact and Gaze Awareness in Video

Anne Kathrine Jensen
Department of Computer Science, Aalborg University
akje08@student.aau.dk

Thomas Søndersø Nielsen
Department of Computer Science, Aalborg University
tsni08@student.aau.dk

Jacob Haubach Smedegård
Department of Computer Science, Aalborg University
jhsm06@student.aau.dk

ABSTRACT

In traditional video mediated communication, eye contact is unachievable due to the offset between the display surface and the camera. We present a novel approach to enabling eye-contact in a video mediated setting, called EyeGaze, which is able to construct a virtual representation of the environment in real-time and present a rendered image from a virtual camera angle that would be impossible in a real world environment. We take inspiration from the volumetric voxel representation of KinectFusion and show how the concepts can be applied to a video conferencing setting. This allows us to present a model with reduced noise and strong resilience to missing data. We frame our solution in relation to previous attempts at enabling eye-contact in video and describe in detail the concept of EyeGaze and the practical implementational details of our pipeline. We then provide a formative assessment of the resulting image quality and the system performance. Finally, we discuss the results and provide a conclusion along with possible directions for future work.

INTRODUCTION

Eye contact is an important communicative tool in face-to-face conversations between humans. We use eye contact to provide information, regulate interaction, express intimacy and exercise social control in social and professional settings [8]. Video conferencing systems allow us to project our image and voice over great distances to facilitate communication without having a physical presence. However, the majority of contemporary systems are unable to support eye contact, and many other non-verbal cues. In an ever more globalised world, communication is often done through different technological solutions. It is therefore problematic that our current solutions in this field do not offer a broader range of human interaction capabilities, such as eye contact.

To establish eye contact, both participants must be simultaneously looking at the each other's eyes. Stokes [16] has shown that at a normal conversational distance, the mutual viewing angle must be between 0-5 degrees for participants to feel they achieve eye contact. This is difficult to achieve using a video conferencing system, as the participant must look at the screen to see their conversation partner's eyes and at the same time look at the camera. This would require the camera to be positioned in front of the screen, obstructing the user's view of their conversation partner, or behind the screen, which would, which would render the camera useless with contemporary opaque screen technologies. For this reason, contemporary screen and camera technology typically place the camera at the edge of the screen, and thus both participants cannot achieve a sense of eye contact at the same time.

The camera positioning problem is a fundamentally challenging issue. Solutions to this problem ranges from strict management of the environment ([11], [12], [14]) to computational solutions based on video or 3D scanning data ([5], [9], [13], [19], [20], [21]). Strict management of the environment, such as camera position in relation to the screen and the user can often create the illusion of eye contact by controlling the distance, and thereby the angle, between the camera and screen from the position of the user. Such systems are often vulnerable to small changes or unintended use of the system, breaking the intended effect, such as users sitting incorrectly, or more users using the system than it was designed for. Alternatively, video and 3D scanning solutions offer adaptive solutions to establishing eye contact, by modifying the video feed, or reconstructing the remote user virtually. These solutions face challenging problems, such as realistically recreating users, and filling out missing data, which the camera cannot see but the modified viewing angle reveals. The arrival of Microsoft's Kinect sensor, which has a built-in infrared 3D-scanner and RGB camera, allows for a cheap alternative to some of the existing solutions while presenting a clean interface for managing both video, depth, and skeleton data.

In this paper we present a method for using multiple Kinect cameras to create a single real-time 3D model of a dynamic environment that is able to remember details of the environment even when they are occluded from view. We begin by exploring existing systems and their basic concepts. We then present the concept behind the EyeGaze system and explain the technology and technique used to capture and update the scene in real time from multiple camera sources. We show how we can create a highly parallelised ray casting engine that can render the scene from any virtual viewpoint within the environment. Furthermore, we use the built-in skeletal data for the user to place the virtual camera based on the observers position, allowing for a natural viewpoint into the remote scene. Finally, we discuss the quality and performance of our system and suggest future work.

RELATED WORK

In our research we investigate the use of eye contact and face gaze. We also investigate other video conferencing solutions, which can be catagorised into blended interaction spaces, and virtual viewpoints.

Eye Contact and Face Gaze

There has been much research within the area of eye contact and face gaze since the 1960s. Kleinke [8] summarised much of this work in his detailed review. Kleinke found that face gaze contributes significantly in the areas of *a*) judgements of liking and attraction, as well as attentiveness, competence, social skills, credibility, and dominance; *b*) important non-verbal regulatory functions in interpersonal communication, such as turn-taking and synchronisation; *c*) social control, for acts such as persuasion, deception, ingratiation, dominance, and compliance; *d*) and finally in "service-tasks" such as teaching and facilitating communication. Clearly, face gaze and eye contact are large parts of communication between individuals, and their absence can lead to an altered perception of people, as well as conversational breakdowns.

Early work on video collaboration

The problem of creating a sense of presence in video communication used by geographically distributed teams has been the focus of research since the mid 80's at Xerox PARC Palo Alto, Portland and EuroParc [6] [17] [1]. Hydra is one of the research projects developed at Xerox PARC during the period [4]. Unlike the more traditional big screen video conferencing, Hydra focused on a four-way roundtable meeting with support for gaze cues, head turning, gaze awareness, and turn taking. This was done by representing each participant using a video surrogate (Hydra unit) consisting of a very small screen with a video camera placed directly under it and a loudspeaker. The distinct placement of these units in a semi circle around each participant, allows for understanding of gaze direction and allows for orientation using sound. Having distinct microphones and speakers for each participant also allows for what Sellen describes as the "cocktail party effect" allowing multiple conversations to be held simultaneously.

Blended Interaction Spaces

Blended interaction spaces describe a group of solutions which use physical placement of displays, cameras, and furniture to create the experience of two or more remote locations blending into one. The idea is to create a solution that facilitates gestures, such as pointing, and, to some degree, face gaze.

Two video conferencing systems based on similar design principles are HP Halo and BISi. HP Halo was analysed by O'Hara et al. [12] as a commercial blended space video conferencing system. BISi was developed by Paay et al. [14] based on their analysis of HP halo. Both setups consists of two identical rooms containing a long curved tabled divided into sections for seating participants. Large displays are mounted on the wall with high resolution cameras mounted above them. When seated correctly, HP Halo provides a life-like and -size display of the remote location, as if the remote location was physically sitting on the other side of the table. HP Halo does not support eye contact, due to the large offset between camera and display, however it does support spatial gesturing, such as pointing and face gaze. Paay et al. found that the optimal camera view for their setup, BISi, would be one or more camera views positioned several metres behind the screens. While this is not physically possible, they speculate that such a view could be achieved with virtual cameras, however this is never tested in their BISi protoype.

Nguyen et al. [11] developed an alternative solution for creating a blended interaction space. Their solution, MultiView, takes advantage of a retroreflective screen to project an image specific to the viewing angle. This allows multiple users to sit in conference on the same screen, but each perceive a different image, relative to their position. MultiView makes use of one camera and one projector per participant to create the unique view points, as well as a retroreflective screen per remote location. MultiView offers a solution to separating the video output individually to each user, but it does not solve the issue of the angle disparity between the screen and the camera.

Virtual Viewpoints

Unlike blended interaction spaces, where furniture, camera, screen, and user positioning are essential to creating an experience of a remote space blending into the local physical space, research in computer vision and 3D graphics has instead focused on manipulating the camera video to create a video conferencing solution. This research is primarily divided into two categories, *1*) RGB cameras generating coloured bitmaps for each frame, and *2*) depth sensing cameras generating bitmaps of registered depth-from-camera each frame.

RGB Cameras

Research making use of RGB cameras can be split into two groups: using two cameras in a stereoscopic analysis to generate 3D coordinates, and texture mapping approaches. An early example of using stereoscopic 3D is Ott et al. [13], which calculated a virtual viewpoint based on stereoscopic analysis between two views. This was done by calculating the pixel correspondance between the views, and then rotating one of the views according to the generated disparity map.

Examples using the texture mapping approach have an intermediate step of generating a 3D model of the user's head, which can be textured using the RGB output of the camera. Yoon and Lee [20] take this approach, using an ellipsoid head model, citing the need for a computationally fast algorithm. Alternatively, Yang and Zhang [19] make use of a personalised head model to improve the quality of the texturised model.

Gemmel et al. [5] developed a system that uses a hybrid model. A virtual avatar is combined with a predefined facial model of the user. The facial model is a rough approximation of generic facial features, and is equipped with synthetic eyes which are rendered on the basis of the video feed. The rest of the model is simply texturised from the video feed. The goal of this system is to better facilitate the experience of eye contact.

Depth Sensing Cameras

Zhu et al. [21] make use of a single SwissRanger SR300 Time-of-Flight depth sensor in conjunction with three RGB cameras. Their solution combines stereo-matching between two RGB cameras with the data from the depth sensor, allowing them to generate a 3D point cloud of the user. After generating a 3D model from the pointcloud, all three RGB cameras are used for texturing.

So far, the discussed virtual camera solutions focus on rendering the head of a person, and not the body, which is necessary for gestures such as pointing. Maimone and Fuchs [9] present a system that captures the entire physical environment in real-time 3D, and then render a personalised viewpoint on 3D screens. 3D capture is achieved using five Microsofts Kinect sensors positioned strategically to capture both foreground and background. The system uses a frame-by-frame model generation algorithm, and does not retain data from previous frames to build the new model. Hole-filling is performed on each individual Kinect, using a modified median filter, without taking into account other Kinects. The system also does colour matching for colour images between Kinects, and 3D eye tracking. Missing edges however, are not recreated or refined over time, as the system uses only the data available in that frame.

EYEGAZE

Our approach to enabling eye contact between two conversation partners in a video mediated conversation is focussed on recreating the natural face to face conversation known from everyday life. Instead of augmenting the raw RGB video feed captured by the cameras, we tackle the deficiencies of the traditional video mediated system by using a virtual camera in a real-time 3D reconstruction of the environment. This allows us to render video from any angle we need, including from positions which would otherwise be impossible, such as camera placements behind the screen.

This also enable us to handle almost any arbitrarily shaped object or person. Like the solution presented by Maimone and Fuchs [9] we make use of commodity depth sensing cameras placed strategically around the display area, such that each camera is able to capture parts of the environment that are occluded from the views of the other cameras. One of the strengths of our solution is the loose relationship between the virtual camera position and the physical camera placement. Since we render the scene from a virtual viewpoint the placement of the physical camera simply becomes a question of scene coverage rather than whether the position results in the correct angle for eye contact. We focus on capturing a single person within the environment due to limitations in the display technology when scaling to multiple persons.

Decoupling the virtual camera angle and the physical camera angles mean that we can construct a similar to the setup illustrated in Figure 1. This setup consists of a large flatscreen display surface and three Kinect sensors. The Kinects are placed strategically on top of of and to the sides of the screen to cover as much of the user as possible. This allows us to capture both sides of the head and detailed facial features as

Figure 1: An EyeGaze setup, using a large flatscreen TV and three Kinects

depth data which can be combined with colour video to produce a fully textured head, with the obvious exception of the back side of the head and neck. We use a single voxel grid to represent the model, by merging the point clouds from all three cameras into the same grid using known extrinsic values for each camera and a projection formula provided by Microsoft. Rendering this model from the virtual viewpoint behind the screen allows us to facilitate eye contact, as can be seen in Figure 2. By utilising head tracking, eye contact can be maintained as the users move from side to side.

Microsoft Kinect for Windows

Our system makes use of the Microsoft Kinect for Windows for scene construction and visualisation. The Kinect sensor is able to produce depth images of an environment by projecting infrared light into the scene and then capture it again using an infrared camera. It does this by using a structured light technique, and can reliably provide depth readings within a distance of 0.4 to 4 metres with near mode activated. Because each Kinect camera uses a similar structured light pattern to calculate depth frames, multiple overlapping Kinects may interfere with each others' patterns. Maimone and Fuchs [9] studied the interference, and found that the Kinects reported no data, rather than faulty data when interference prevented depth readings. The infrared camera's field of view is approximately $45.6°$ vertically, and $58.5°$ horizontally with a maximum depth frame resolution of 640×480 capturing at a frame rate of 30 frames per second. This allows for smooth real-time updates.

The Kinect is also able to capture a traditional colour video feed. The colour camera can capture frames at a resolution up to 1280×720, though at a diminished frame rate of only 12 frames per second. To avoid a frame rate bottleneck stemming from the Kinect, we make use of the 640×480 colour resolution, which, like the depth camera, provides a frame rate of 30 frames per second. The field of view for the colour camera is approximately $48.6°$ vertically and $62.0°$ horizontally, which is different from the depth camera, however, the official Microsoft Kinect SDK provides functionality to map the colour image to the depth map using internal intrinsic and extrinsic parameters.

The official Microsoft Kinect SDK also provides skeleton

Figure 2: A comparison between the 3D models and their texturised version

tracking functionality for the Kinect sensor. The SDK allows for both seated and standing tracking of up to 20 joints, and can track up to two skeletons simultaniously.

IMPLEMENTATION

In conjunction with the Microsoft Kinect sensors, our implementation relies on the DirectX framework for a fast GPU based implementation. The advances in GPU programming allows us to offload large highly parallelisable tasks to the GPU, such as updating data within a large 3D data structure (a voxel grid) and ray casting the data to produce an output image.

Our graphics pipeline (as shown in Figure 3) can be divided into two main parts responsible for input and output, respectively. The first main part of our pipeline is the merging algorithm responsible for capturing depth, RGB, and skeletal data from the Kinect sensors. Skeletal data is passed to the network interface and the depth data to the GPU. On the GPU the depth data is converted into 3D points in world space and merged into the voxel grid, stored in GPU memory.

The second part is a rendering engine responsible for rendering the final output image for display on the display device. This part of the pipeline is responsible for calculating the correspondence map between the RGB data and the depth data, receiving skeletal data via a network interface, and passing the RGB and skeletal data to the GPU. The skeletal data is used to decide the position of the virtual camera, and the textures are used when ray casting the voxel grid.

Before we describe implementation details for our pipeline, we describe the technique we use to calibrate the cameras, and calculate their extrinsic values.

Calibration

Using multiple cameras for optimal scene coverage requires exact knowledge of the placements of the cameras in relation to each other. Without precise extrinsic camera values for each Kinect, the depth and colour frames will not overlap for each Kinect, causing holes, scars, and bumps to appear in the model, or in extreme cases, generating the model twice, as if the user is experiencing double vision. To calculate the correct placement of the cameras, we calibrate each camera against a common real world target object visible by each Kinect. We chose a 13 x 9 grid checkerboard pattern which provides us with 117 corner points for which we can obtain approximate depth values. Using the checkerboard recognition algorithms provided by the OpenCV library we can obtain the pixel coordinates in the RGB frame for each corner point. Given the physical distance between the RGB sensor and the Infrared sensor in the Kinect sensor, we map the colour frame to the depth frame using the Kinect API provided by Microsoft, which allows us to do a direct mapping between colour pixel coordinates and depth values which in turn can be mapped to 3D space by using the Kinect API.

Having obtained the 3D point for each internal corner on the checkerboard from each Kinect, we use an ICP algorithm for calculating the transformation matrix (rotation matrix and translation vector) between each camera. For this purpose we use an ICP algorithm for Matlab [18] to run ICP on the point clouds of the checkerboard corners for each Kinect. For each flanking Kinect we calculate the transformation matrix from the Kinect above the screen to it, thereby allowing us to automatically place all Kinects relative to a fixed known position for a single Kinect.

During our research we noticed a discrepancy between the 3D points obtained for each internal checkerboard corner using the Kinect API and the actual real-world distance between these points, which results in an offset after running the ICP algorithm. We corrected this by manually fine-tuning the translation after the calibration step.

Volumetric Representation and Scene Reconstruction

Our goal is to render the scene as close to reality as possible. Taking into account the uncertainty of the Kinect sensor depth values, we draw upon the work of Izadi et al. [7] who

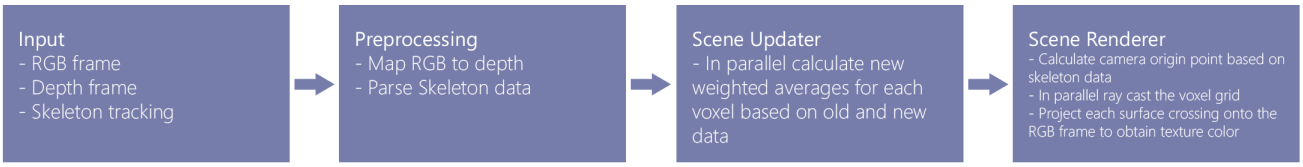


Figure 3: Pipeline for our implementation, describing the major components of our implementation

made use of a voxel grid representation of a 3D model via a truncated signed distance function (TSDF). Voxels are in effect volumetric pixels, which instead of containing colour information, contain the distance from the voxel to the nearest surface. Unlike polygons, voxel positions are implicit and relative to each other, which means that the amount of information we need to store is significantly reduced. The implicit data structure and the fact that it stores relative distance values makes voxels ideal not only for discretely representing a TSDF but also for storing a 3D model based on the uncertain depth values provided by Kinect sensors.

A signed distance function (SDF) was used by Curless et al. [2] as a volumetric method for building complex models from range images. The SDF maps 3D coordinates to distance values to a nearby surface. The distance values are zero at the surface crossing and assume postive values further away from the surface. Passing through the surface and thus being within the object results in negative distance values. The TSDF used by Izadi et al. [7] to store a scanned object modifies the SDF by truncating the maximum surface distance a voxel can store to a value based on the uncertainty of the scanned depth values. Truncating the distance values can cause some surfaces not to be represented, for example, when the surface runs close to parallel with the camera view direction, however, in practice this rarely occurs.

In our implementation, the TSDF is implemented as a large single-dimensional array, stored in GPU memory. We use a fixed voxel grid size of $512 \times 512 \times 512$, using 4 bytes per voxel. This is not memory efficient, consuming 512MB of memory, however, it allows quick and easy memory access. Due to its large size, the TSDF exists solely in GPU memory, and is never transferred to main memory or accessed by the CPU, as this would require excessive memory transfers and read/writes. Since all TSDF updates and operations are performed in parallel on the GPU, we leverage the highly parallel nature of contemporary GPU programming to achieve real-time interactive frame rates.

Scene Updates

Once new depth frames from the Kinects are received, they are transferred to the GPU, where they are merged into the voxel grid representation of the scene. Each voxel in the voxel representation is independent from each other, meaning that we can update them independently in parallel. During a scene update, the GPU steps through the voxel grid in XY slices on the Z-axis, updating each slice with new values from the new depth frames, for each frame. To to this we create 512×512 GPU threads to perform the scene update calculations, where each thread is responsible for updating a single row of voxels on the Z axis.

Algorithm 1 Merging algorithm for merging multiple Kinect depth frames into a voxel grid

Require: Weighting values $\text{weight}_1, \text{weight}_2$ and a truncation value truncval

```
1  function SceneUpdate(TSDF, cameras, dim)
2      for each voxel g in x, y slice of TSDF in parallel do
3          for depth d = 0 to dim do
4              g_dist ← TSDF(g)
5              g'_dist ← 0
6              v_g ← convert g_d to world space
7              for each Kinect C in cameras do
8                  v ← transform v_g into camera space
9                  p ← project v onto the depth frame
10                 dist ← Frame(C, p) − v
11                 if p is within the depth frame & Frame(C, p) is within the scene & dist < −5 then
12                     g'_dist ← g'_dist + dist
13                 end if
14             end for
15             g'_dist ← average g'_dist over cameras
16             g'_dist ← (g_dist · weight_1 + g'_dist · weight_2) / (weight_1 + weight_2)
17             if g'_dist > 0 then
18                 g'_dist ← min(truncval, g'_dist)
19             else
20                 g'_dist ← max(-truncval, g'_dist)
21             end if
22             TSDF(g) ← g'_dist
23         end for
24     end for
25 end function
```

Algorithm 1 efficiently processes the voxel grid by dividing it into slices on the Z-axis. A GPU thread is assigned to each voxel in the slice and processes the specific XY voxel in each slice sweeping along the Z-axis. Each thread must take into consideration data from each connected Kinect. In Line 8 we convert the world space coordinates of the voxel to camera space, by using the transformation matrix for the specific camera we are investigating. We then project the camera coordinates to the depth frame in Line 9. The Kinect SDK provides a number of methods for converting between colour, depth, and 3D coordinates, however the exact underlying intrinsic and extrinsic values for the colour and the infrared cameras are not publicly available. The API is also not available on the GPU. In order to convert from 3D points in

camera space to depth image points using the GPU, we rely on the constants provided online by Microsoft [3] which are a rough estimate of the real values for the specific Kinects. An alternative is to use a standard perspective projection matrix based on the field-of-view of the infrared camera, however this produces a warped 3D point cloud with erroneous values, suggesting that the depth frame data is not consistent with the infrared camera output.

Using the depth frame coordinates, we can perform a lookup in the depth frame to determine the distance from the sensor plane to the surface. We subtract this distance from the distance of the voxel from the sensor plane in Line 10 to find the distance from the voxel to the surface. If this distance is negative beyond a threshold, then the voxel is behind the surface, from the perspective of the camera, and this camera is ignored.

Once a new distance value has been found, it is merged into the TSDF in Line 16, by means of a weighted average. By weighting the previous data higher, we can control how sensitive EyeGaze is to sudden changes in the depth frame, such as a sudden, temporary loss of data on a depth frame pixel. This is a trade off between image stability and update speed, as changes in the real world will reflect slower in the virtual reconstruction the greater we weigh previous data. Finally, the last part of Algorithm 1 truncates the new values if they exceed the truncation value.

When working on the GPU, it is important to reduce branching, as the GPU is not as capable of branch prediction as the CPU. There are performance penalties especially when individual threads can execute separate branches simultaniously on the GPU. The GPU is capable of simulating branching by executing all branches and deciding which branch was correct; however, this is potentially very costly. Algorithm 1 has been designed to minimize branching in order to improve performance.

Rendering the Scene

As described in the previous section, the 3D representation of the environment is stored within the TSDF as an isosurface. Rendering the scene can be done in two ways, *a*) generate polygons using the marching cubes algorithm, and render using standard rasterisation techniques, or *b*) ray cast the voxel grid by determining the surface zero-crossing.

Generating a polygon mesh would allow us to utilise high speed rasterisation techniques, however, given the widespread availability of powerful GPU's, we implemented a GPU-accelerated ray caster which greatly simplifies the pipeline. Furthermore, this allows us to minimise any potential loss of detail that could occur from the transformation of the isosurface to polygons.

Typical rendering solutions using virtual cameras make use of symmetric camera frustums, where the distance from the camera position to each corner of the viewport remains the same. Moving the camera position with this type of camera also moves the viewport. This is however not the case for an asymmetric view frustum, where the camera position can move independently of the viewport.

EyeGaze makes use of the asymmetric view frustum. In EyeGaze, the viewport is stationary, and acts as a window into the remote location. The camera position is a virtual representation of the user's eyes, and moves around with the user, allowing him to look around corners and replicating the effect of looking through a window into an other room. In EyeGaze, the viewport is positioned on the edge of the voxel grid, allowing the entire voxel grid to be used for remote location capture. The viewport's physical dimensions are sized to match the screen the user is looking through. We calculate the origin r_o of the ray from the 3D position of each pixel on the viewport. Using the origin and the camera position p, we can calculate the ray direction $r_d = r_o - p$.

In order to estimate the surface zero-crossing, we use the method of triliniear interpolation, presented by Parker et al. [15]. In their paper, Parker et al. present a method for ray casting isosurfaces at interactive rates by using trillinear interpolation to estimate the ray's current position within the voxel. This method also allows the surface normal to easily be estimated. Using this method, we march rays from the viewport through the voxel grid until the ray reaches a voxel in which a zero-crossing exists. Since all TSDF values have been truncated to a maximum truncation value s, we know that once a voxel v yields a TSDF value $TSDF(v) < s$, a surface zero-crossing is within this voxel, or nearby. Trilinear interpolation is then used to locate the zero-crossing by calculating the position on either side of the zero-crossing. These two points can be used to calculate the zero-crossing, by calculating the ray length t^* given the formula

$$t^* = t - \frac{\Delta t F_t^+}{F_{t+\Delta t}^+ - F_t^+}$$

where t and $t + \Delta t$ are ray lengths to points on either side of the zero-crossing, and F_t^+ and $F_{t+\Delta t}^+$ are interpolated distances on either side of the zero crossing.

Since we intend to texturise our model using the RGB camera, we do not need to bounce the rays off the intersected surface for light or shadow calculations. This gives us a great advantage in terms of performance, as our ray caster only needs to intersect a surface once for each ray.

QUALITY AND PERFORMANCE

We will now present the different quality improvements that our system entails and compare the results to a straightforward implementation using a single Kinect and depth frame rasterisation. In this section we present results on eye contact, render output, interference, model quality, environment retention, and head tracking.

Render Quality

EyeGaze is able to capture a person and his or her immediate environment from three distinct viewpoints in 3D and merge these viewpoints into a single, global 3D representation. Figure 4 shows three images from three different systems (*a*) the raw RGB camera feed, (*b*) a simple frame-to-frame 3D model generated by quads created from depth points and texturised using the RGB camera, and (*c*) our EyeGaze solution, using three Kinect sensors to generate a complex and detailed 3D

(a) The raw image from the camera, capturing the person from a downward perspective.

(b) Image captured by a simple method of 3D generation, where quads are created from depth points, and texturised

(c) Output image from EyeGaze, texturised using the RGB video of a single camera

(d) Output image from EyeGaze using two cameras for texturing

Figure 4: Screen captures from three different systems

model. Missing data is represented as the magenta colour in Figures 4b and 4c. As seen, there is significantly less missing data in EyeGaze, as the side cameras are capable of filling in missing information. In addition, EyeGaze is much less sensitive to sudden holes in the depth frame as previous data is simply retained, which would be instantly visible with the method used in Figure 4b. This gives a cleaner image, with less flickering in Figure 4c.

The texture resolution and quality of EyeGaze is dependent on the RGB camera used. In this implementation we rely solely on the built-in RGB camera of the Kinect sensor. The RGB camera of the Kinect sensor produces a low quality output image, which is even lower than most stand alone webcams available today. At the same time, the Kinect RGB camera has a wide field of view which means that the percentage of the low quality RGB frame that is used to texture the person is fairly low. This lack in quality results in a blurry output render image when mapped to our life-sized model, degrading the overall quality.

Texture mapping approaches

We have explored two approaches to texture mapping the generated model with the Kinect RGB video. The currently used approach is to use a single Kinect which has a view of the majority of the surface area. In our setup, we use a Kinect positioned on top of the screen and tilted slightly downwards. This method, however, limits the areas of the model for which we have live texture. Given its placement, the Kinect cannot see some areas underneath the chin, which causes EyeGaze to replicate the chin texture on the area below the chin. In order to solve this issue another viable solution is to use the calculated surface normal vector and compare it to the direction vector of the Kinects. By comparing the angles between the Kinects and the normal vector, we can determine which Kinect has the most head-on view of the surface at that specific point. We can further enhance this by performing a depth check on the depth frame from that Kinect, to ensure that nothing occludes the view of the surface from that Kinect. This solution would in theory give the best possible texture to each area of the surface, however, this solution still requires more work as different light conditions and slightly different camera settings cause the texture to appear strange and unnatural when mixed from different Kinects. Furthermore, it requires exact knowledge of the field of view of each depth camera, RGB camera, and the exact parameters for converting depth points to 3D points. An example of the split texture can be seen in Figure 4d. It is worth noting how this approach already shows some clear advantages over the single texture approach of Figure 4c. Especially looking at the person's ears and sides of the head, which are much more well defined than when simply using a single Kinect. However, the image quality suffers where the two textures overlap, since the model here is a merger of two sligtly different models with respect to the depth to 3D conversion.

Model Quality

In Figure 5 we show the model resulting from merging the Kinect data from two Kinects into a single coherent model. What is worth noting is that we are able to represent facial

Figure 5: Each column depicts the image and model being generated from each Kinect on either side of the screen. The bottom image is the resulting model when the two individual but incomplete models are merged

features in great detail. We can represent the receeding eye sockets, the extruding nose, and even cheekbones. Furthermore, we are able to render the front and sides of the torso and arms thereby supporting transfer of gestures and body language across the video feed in addition to eye gaze direction. Unlike other attempts using approximations to the body shape and facial structure, our solution is capable of creating an accurate detailed personal model each frame that can be textured as needed.

Looking at Figure 4 we see how we are able to render the scene from a virtual camera angle behind our display surface. This angle allows EyeGaze to facilitate eye-contact. Furthermore, eye-contact is maintained through the use of head tracking, allowing the virtual camera to mimic the movement of the user.

Adding to the model quality is the model retention properties of the voxel representation. Since we only update voxels in front of, and immediately behind, the object surface, we are able to retain detail which would be occluded from view due to the dynamic nature of the environment.

Interference

The Kinect sensor was designed to be used as a single device covering a space using a structured light IR projector. Using multiple Kinects to cover the same area results in some interference between the IR patterns projected by the different Kinects. The IR interference takes several different forms. The first distinction between the interference is between wrong depth readings and missing data. According to Maimone and Fuchs, the majority of the interference is represented as missing data. We found that the majority of the missing data interference is sporadic, thereby allowing for depth readings in some frames which our model is capable of handling. However, we also noted that when using more than two Kinect sensors the amount of permanent interference rose, which meant that some parts of the environment and persons within it would never be rendered. This suggests that there is a sharp diminishing return when adding additional Kinects to cover the scene from a new angle.

Performance

We ran our system on Intel Core i7 Ivy Bridge, with a substantial amount of memory and a NVIDIA Geforce GTX 560 graphics card. Comparing the performance data in Table 1 we see decreases in frame rate occuring when the resolution is increased, and additional cameras are added. The frame rate decrease is greater when adding more Kinects, than when increasing the resolution. This can be seen when comparing the frame rate decrease from 720p to 1080p between one and three Kinect cameras. As we use more cameras, the frame rate impact of increasing the resolution because less severe. This indicates that the majority of the computation time is spent on merging depth frames into the TSDF, which is independent of the screen resolution.

Kinects	300×300	480p	720p	1080p	3200×1800
1	30.0	28.3	21.5	13.8	7.7
3	14.0	12.7	11.5	9.5	6.0

Table 1: FPS measured for different resolutions, using a different number of Kinects. Merging performance can be compared vertically and ray caster performance can be obtained by comparing numbers vertically

At the resolution of 300×300 the implementation using a single Kinect is able to run at maximum Kinect frame rate, but using three Kinects results in a performance loss of 53%. Running at 720p with three Kinects still provides interactive rates, however the experience becomes unsatisfactory when the resolution is increased to 1080p or higher, resulting in reduced sense of presence [10]. It is worth noting the performance of our ray caster, which is able to render the scene efficiently within miliseconds even at 720p. This is due the

the parallel implementation relying on the number of cores on the GPU.

DISCUSSION

Comparing the quality of the model with the quality of the final rendered image shows that some work is still needed to make the experience natural. For simplicity we have focussed only on rendering a single texture on the model from the camera placed above the screen. This results in some texture tear when the difference between the surface normal and the camera direction becomes too extreme. In addition, there are some inaccuracies in the calculations converting from depth pixels to 3D points in camera space, and vice-versa. This means that when we perform matching between the projected 3D coordinates and the colour coordinates of a specific sensor, the parts of the model which have been generated by merging data from other sensors are slightly misaligned with the texture. In order for a better experience, a closer analysis of the 3D point cloud has to be made in order to obtain the correct values for the depth to 3D point conversion. This would allow for use of any arbitrary texture, and a better mapping between model and texture.

Using the default skeletal tracking system of the Kinect sensor has some benefits in terms of performance and some disadvantages in terms of the stability of the skeleton. The Kinect skeletal tracking system is stable most of the time, but suffers from edge cases where the skeletal engine is unsure about the exact position of the skeleton, making the skeleton points move erratically. This causes the rendered image to jump around for the user, as the virtual camera position is determined from the tracked position of the user's head. An alternative approach would be to use the facial recognition of the OpenCV library to do eye tracking on the RGB frame, which would eliminate some of the challenges of using the Kinect skeletal tracking.

In our implementation we have not focussed on supporting communication such as sound interfaces or spatial layout of the room and lighting. These are all aspects which have some implications for the final experience and the immersion into the experience. Especially the synchronisation between sound and image can become a problem.

We have shown that we are able to render the scene at interactive frame rates using the current mid-range graphics hardware of our setup. There are several possible optimisation paths. The most straightforward approach is to upgrade the graphics card, allowing EyeGaze to parallelise to more shader cores, and perform the computations more rapidly. In the implementation, updating the voxel grid and rendering is currently in lockstep, which means that for each update there is a single rendering. This limits the frame rate of the output and removes the possibility of updating the scene from a new viewpoint before the next update. A better approach would be to decouple the updating and the ray casting of the scene.

A limitation of the system is the sometimes poor rendering quality of individuals wearing glasses, as the depth sensor has difficulty registering glasses due to their small and reflective surface area. Similarly reflective or transparent objects present a challenge for the system as the sensor is unable to register the surface.

CONCLUSION

We have presented an alternative approach to video conferencing systems that enable eye contact and gaze awareness using the Microsoft Kinect sensor. Our novel system, EyeGaze, captures the user's envinronment in 3D in real-time, allowing a virtual camera to be positioned wherever provides the best experience of eye contact. Furthermore, EyeGaze tracks the user's head to continually provide the correct virtual camera placement. This also allows EyeGaze to give the effect of looking around corners, through the use of an asymmetric virtual camera. Through the use of a volumetric representation, our system can merge the output from multiple Kinects into a single model, which captures more details of the environment than a single Kinect and allows for a robustness against sudden gaps or erroneous data. We have framed this implementation in regards to related work within video conferencing, specifically Blended Interaction Spaces, virtual camera implementations, and the groundwork laid by Buxton and Sellen in gaze aware video communication. Furthermore, we have provided implementational details of our solution and provided a formative assessment of the model and render quality, and the performance of the implementation. Lastly, we have provided a discussion of the results and pointed towards future improvements in the system.

Future Work

There are several improvements to EyeGaze which we would like to explore in future research. Firstly, although using multiple Kinects allows us to generate a more complete 3D model, texturing the 3D model only uses one Kinect for now. Using split texture still requires some work before it is a viable solution.

EyeGaze currently only supports two users, however, this is a limitation we have enforced due to limitations in screen technology. EyeGaze is capable of rendering any number of potential viewpoints, given capable hardware, for any number of potential users. This feature could be used both to facilitate additional users, but also to facilitate 3D screens. The primary issue in supporting multiple users is finding a screen technology that can deliver each user their own viewpoint. One possible solution to this is taking inspiration from Nguyen et al. [11], and using projectors along with a retroreflective screen surface.

REFERENCES

1. Buxton, B., and Moran, T. Europarc's integrated interactive intermedia facility (iiif): early experiences. In *Proceedings of the IFIP WG 8.4 confernece on Multi-user interfaces and applications*, Elsevier North-Holland, Inc. (Amsterdam, The Netherlands, The Netherlands, 1990), 11–34.
2. Curless, B., and Levoy, M. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, SIGGRAPH '96, ACM (New York, NY, USA, 1996), 303–312.

3. Elsbree, J. http://social.msdn.microsoft.com/Forums/en-US/kinectsdknuiapi/thread/e53a4ba7-2522-407f-9d60-86e6fc5f89dc/.

4. Finn, K. E. *Video-Mediated Communication*. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 1997.

5. Gemmell, J., Toyama, K., Zitnick, C. L., Kang, T., and Seitz, S. Gaze awareness for video-conferencing: A software approach. *IEEE MultiMedia 7*, 4 (Oct. 2000), 26–35.

6. Goodman, G. O., and Abel, M. J. Collaboration research in scl. In *Proceedings of the 1986 ACM conference on Computer-supported cooperative work*, CSCW '86, ACM (New York, NY, USA, 1986), 246–251.

7. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., and Fitzgibbon, A. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST '11, ACM (New York, NY, USA, 2011), 559–568.

8. Kleinke, C. L. Gaze and Eye Contact: A Research Review. *Psychological Bulletin 100*, 1 (1986), 78–100.

9. Maimone, A., and Fuchs, H. Encumbrance-free telepresence system with real-time 3d capture and display using commodity depth cameras. In *Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality*, ISMAR '11, IEEE Computer Society (Washington, DC, USA, 2011), 137–146.

10. Meehan, M. *Physiological Reaction as an Objective Measure of Presence in Virtual Environments*. PhD thesis, University of North Carolina at Chapel Hill.

11. Nguyen, D., and Canny, J. Multiview: spatially faithful group video conferencing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, ACM (New York, NY, USA, 2005), 799–808.

12. O'Hara, K., Kjeldskov, J., and Paay, J. Blended interaction spaces for distributed team collaboration. *ACM Trans. Comput.-Hum. Interact. 18*, 1 (May 2011), 3:1–3:28.

13. Ott, M., Lewis, J. P., and Cox, I. Teleconferencing eye contract using a virtual camera. In *INTERACT '93 and CHI '93 Conference Companion on Human Factors in Computing Systems*, CHI '93, ACM (New York, NY, USA, 1993), 109–110.

14. Paay, J., Kjeldskov, J., and O'Hara, K. Bisi: a blended interaction space. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, ACM (New York, NY, USA, 2011), 185–200.

15. Parker, S., Shirley, P., Livnat, Y., Hansen, C., and Sloan, P.-P. Interactive ray tracing for isosurface rendering. In *Proceedings of the conference on Visualization '98*, VIS '98, IEEE Computer Society Press (Los Alamitos, CA, USA, 1998), 233–238.

16. Stokes, R. Human factors and appearance design considerations of the mod ii picturephone amp;#174; station set. *Communication Technology, IEEE Transactions on 17*, 2 (april 1969), 318 –323.

17. Stults, B. Media space. Tech. rep., Xerox Parc, 1986.

18. Wilm, J. http://www.mathworks.com/matlabcentral/fileexchange/27804-iterative-closest-point.

19. Yang, R., and Zhang, Z. Eye gaze correction with stereovision for video-teleconferencing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 26*, 7 (july 2004), 956 –960.

20. Yoon, N.-R., and Lee, B.-U. Viewpoint interpolation using an ellipsoid head model for video teleconferencing. In *Advances in Visual Computing*, G. Bebis, R. Boyle, D. Koracin, and B. Parvin, Eds., vol. 3804 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2005, 287–293.

21. Zhu, J., Yang, R., and Xiang, X. Eye contact in video conference via fusion of time-of-flight depth sensor and stereo. *3D Research 2* (2011), 1–10.

Research Paper 2

EyeGaze: A Comparative Analysis of Video Mediated Eye Contact

Anne Kathrine Jensen
Department of Computer Science, Aalborg University
akje08@student.aau.dk

Thomas Søndersø Nielsen
Department of Computer Science, Aalborg University
tsni08@student.aau.dk

Jacob Haubach Smedegård
Department of Computer Science, Aalborg University
jhsm06@student.aau.dk

ABSTRACT

Eye contact and body language are inherent parts of a natural conversation, and helps interpreting the mood and intentions of a conversation partner. We present a study of EyeGaze, which uses several Microsoft Kinects for building a 3D model of the user and renders it from a custom view point such as to allow for eye contact between two individuals. Related work on eye contact and attempts at achieving this in video communication are presented, and used as a base for a comparative study of the EyeGaze system. The goal is finding differences and similarities in the experience gained when communicating Face-to-Face, via Skype, and via the EyeGaze system. The results reveal that the Face-to-Face condition always gives a better communication experience, and a small, general bias in the favour of EyeGaze when compared to Skype. Responses show a significant difference in the experience of eye contact in EyeGaze's favour when comparing it with Skype. However, in most measures, there is no statistically significant difference between EyeGaze and Skype. We discuss the results and hypothesise on what may have caused the different outcomes.

INTRODUCTION

Eye contact has always been an integral part of human communication. Language (both spoken and body) has primarily evolved under circumstances where communication was only done with people in close vicinity. In this context, eye contact and avoidance of the same has come to carry meaning - and choosing one or the other has been necessary, as speech was most often done with someone in the same room. The invention of the telephone made communication over distances easier, but also removed the option of body language in the communication experience.

As presented by Chris Kleinke [11] in his 1986 review of the field, there is a large body of work within the field of social psychology showing many different meanings of face gaze and eye contact. Face gaze is achieved when the subject looks at the other's face and eye contact is defined as two people simultaneously looking at each other's eyes. It is often indistinguishable for a viewer if a person is engaged in face gaze or eye gaze, and thus measurement of eye contact is often a measurement of mutual face gaze.

Different cultures assign different value to face gaze and eye contact. In some, seeking eye contact conveys respect - and in others, avoiding it is the respectful thing to do. Western research points at eye contact making a difference for how attentive a person seems in a conversation, as well as communicating respect, understanding, attraction and much more. Non-wavering eye gaze can be an expression of intimacy or appear threatening depending on the context and persons involved. This serves as an example of the different meanings inherent in different types of gaze. Furthermore, seeking and breaking eye contact is an often used method for conveying the wish to speak. Thus, eye contact is quite important for ensuring proper turn taking and avoiding interruptions.

Communicating through a medium that does not support body language and gaze can lead to issues pertaining to turn-taking and misunderstandings of intentions. Thus, research has studied how to allow people to see each other for conversations over distance. Two of the most widely known and used tools for seeing the remote partner are Skype and Apple's FaceTime. Due to the design of contemporary camera and screen technology, the camera is most often placed on top of the screen or beside it, rendering it impossible for both participants to have a feeling of eye contact at the same time. For this to be achievable, both participants would simultaneously have to look at the camera, to give the other the impression of eye contact, and at the screen to achieve the feeling for themselves.

In [8] we presented a method for achieving eye contact in a video mediated communication experience by using Microsoft Kinect sensors to build a real-time 3D-model of the user, and then render that person using a virtual camera placed based on the remote user's position in 3D space. This allows both users to look at each other while at the same time giving the other person an impression of being looked at. This approach to rendering makes the screen act like a window into the other room. This type of experience allows for more natural communication which pulls on body language and eye contact for supporting information.

We begin by describing related work within the field of eye contact and face gaze, as well as similar studies on eye contact. We then describe the method used in the study, and

present the findings. Finally, we discuss the findings and ground our observations in the theory of proxemic interaction and conclude with our thoughts on future work.

RELATED WORK

To better understand the effects of eye contact, we explore research performed in the areas of psychology and sociology. Gaze and eye contact have been the subject of much research since the 1960's. In the 1986, Chris Klienke [11] reviewed and aggregated the results of many studies performed, detailing the use of gaze and eye contact in interactions, as well as the effects on the participants. We divide our investigation into three catagories: *(i)* how do people use gaze and eye contact in conversations; *(ii)* what are the effects of having eye contact; and *(iii)* which efforts have been made to facilitate eye contact over video, and how were they evaluated?

How Do People Use Gaze and Eye Contact?

Gaze and eye contact are typically used in social interactions similar to the use of body language. One such example is during turn-taking activities, such as in discussions or conversations where it is appropriate that only one person speaks at a time. In this regard, Kendon [9] found that in two-person interaction, speakers indicated the end of their turn to speak by giving a prolonged gaze directed toward the listener. He also found that the listening party looked away from the speaker to indicate their desire to speak in 70% of the cases involving utterances longer than five seconds. In the same vein, Duncan and Niederehe [2] found that of four various cues indicated to relate to taking the speaking turn (looking away, gestures, breathing indications, and speaking louder) looking away and making gestures were the more effective methods of indicating a desire to speak. Additionally, when using one or more of these cues, problems related to an uncertainty regarding whose turn it was to speak were almost not present.

Kendon notes that subjects' gazing behaviour changed according to whom they were speaking. Thus, a test subject who took part in two dyadic discussions had nearly doubled the mean length of time gazing at the other participant in one pairing from the other - and in both cases, his mean gaze time was within half a second of that of the partner. Furthermore, test subjects' gazing behaviour was different during speaking and listening. Each person had different levels of gazing at and away from their conversation partner, but generally long face gazes were used when listening, and shorter face gazes were used when speaking, as well looking elsewhere.

Gaze and eye contact is also used as a sign of liking and attraction. Kleinke et al. [13] and Thayer and Schiff [18] found that people appear to like each other more when they share high rather than low amounts of gaze.

What Are The Effects of Having Eye Contact?

Having and maintaining eye contact in an interaction can have a significant effect on individuals. The effect tends to be highly situational, and depends on what task the individual is currently engaged in. An individual's degree of gaze and eye contact can have significant influence on perception and ability. This effect is sometimes conscious, and some times unconscious and contradictory to what the users expect from themselves. Fry and Smith [3] found that students performed better at digit encoding tasks when the instructions were read with high amounts of eye contact. In addition to having an effect on attentiveness and performance, Kleinke et al. [12] found that members of the opposite sex judged their conversation partners' attentiveness based on how much they maintained eye contact. Partners who gazed at lower levels were considered to be less attentive, while partners who gazed at higher levels were considered more attentive. Scherer [17] found that test subjects found other subjects more likable when they used higher amounts of eye contact. Kraut and Poe [14] designed a study in which airline passengers were asked to pass through a mock customs check, with fully trained customs officers performing the mock checks. Half the passengers were given a small pouch of white powder to serve as contraband, and a $100 reward was offered to the passenger who could deceive the officers most convincingly. Kraut and Poe found that people appear less trustworthy when attempting to avoid eye contact. Finally, Mehrabian [16] found in his study of attitudes from posture, orientation, and distance of a communicator, that higher levels of eye contact and a more open posture communicated a more comfortable stance. His results are confirmed in Kleinke's [11] review of gaze and eye contact, where high levels of eye contact were found to produce a comforting effect.

Evaluation of Eye Contact

Different methods have been used for evaluating eye contact and meaningful gaze behaviour in interactions between humans and digital avatars. By questioning test subjects through a questionnaire, Garau et al. [4] found that a conversation mediated by avatars whose gaze and head movements were informed by the user's behaviour outperformed both conversations with only audio and conversations in which the avatar had random gazing behaviour and head movement. Their questionnaire had 15 questions subdivided into four groups (face-to-face, involvement, co-presence, and partner-evaluation) and each question was evaluated through a 9-point Likert scale.

Rehm and André [15] tested whether communicative behaviours changed when interacting with an avatar for a computer agent or a co-present human. They let two people and an avatar play a game of dice.The avatar spoke, gestured, and exhibited a gazing behaviour which was based on research into human gazing behaviour during conversations. The test session was recorded, and gazing and speaking behaviour was analysed. Results showed that the subjects' gazing behaviour did not change significantly based on whether they were addressing the other person or the avatar, but that they looked significantly more at the agent than their human partner. This might be because of the novelty of the avatar, or it might be caused by the users not regarding the agent as an equal conversation partner.

Additionally, Kipp and Gebhard [10] performed mock interviews between a user and an avatar for a virtual agent. The avatar was programmed to have a four different types of gazing behaviour. Two were based on research into gazing be-

haviour in contact between humans, and were meant to give the impression of social dominance or submission. The other two continuously looked at the user's head in each of their configuration. Users experienced all four types of gazing behaviour, and answered a questionnaire with five questions after each type. The questionnaire results showed that both the dominating gazing avatar as well as the two types of continuous stare were interpreted as dominant - while the socially submissive behaviour gave the impression of not being dominant, while also being the one which seemed the most extroverted.

Finally, Bee et al. [1] had users take on the role of a character in a story and then speak to an avatar representing another character. They tested two different gazing behaviours of the avatar: an interactive behaviour, based on speech and gaze from the user, and a non-interactive behaviour, in which the avatar appeared to gaze at random. Users rated the experience through a questionnaire. The results indicated that the interactively gazing character gave a better impression on the measures social presence, rapport with the character, engagement, social attraction of the character, and perception of the story.

EYEGAZE

To facilitate eye contact and face gaze in a video mediated conversation, we developed EyeGaze [8], a prototype solution that enables understanding of nonverbal cues such as eye contact and gaze direction over video.

The EyeGaze system uses three Microsoft Kinects, which each have a built-in RGB camera and a depth sensor. The Kinect's depth sensor estimates the depth of objects in front of it by emitting a structured light pattern. The Kinects are placed around the screen to capture as much of the users face and body as possible from various angles. Figure 1 shows one potential EyeGaze setup. EyeGaze generates a persistent 3D-model by continuously merging individual depth frames from the different angles, allowing the model to obtain new details over time. To achieve the sense of eye contact, a virtual camera is then positioned where the conversation partner's eyes would be, rendering an image on the display surface as if the screen was a window. Using the built-in head tracking of the Kinect, the virtual camera is moved accordingly. If the user moves - the virtual camera position moves, thus maintaining the sense that the user is looking through a window. The generated 3D model for this study is texturised using a single Kinect, as using multiple Kinects requires matching the gain, contrast, exposure, brightness and individual color saturations of the RGB frames.

To facilitate this experience the EyeGaze implementation uses the recent advantages of General Purpose GPU programming[7] to store and, in parallel, operate on large data structures. Unlike previous research on using Kinects for video conferencing, our implementation takes inspiration from research within the area of 3D scanning applications using the Kinect. For each Kinect we transfer the captured frame data to the GPU and in parallel merge the data into a single voxel representation of the environment. This voxel grid is stored as a 512mb array, where each voxel is represented as a single float value, with its 3D coordinates designated implicitly by its position in the array. Once merged, we perform a ray casting of the voxel grid, again leveraging the power and parallel nature of the GPU.

Figure 1: Picture showing the EyeGaze setup using a flat screen LCD display surrounded by 3 Kinect sensors. The video rendered on the screen is based on head tracking of the individual positioned in front of the screen.

STUDY METHOD

We wish to understand the implications of facilitation of eye contact in EyeGaze. To this end we have designed a within-subject study of our system compared to two other conditions, namely the traditional video mediated setting and the natural face to face conversation.

Independent Variable

We have three conditions, using different technological means to facilitate the discussion.

EyeGaze
The discussion is facilitated through two EyeGaze setups (Figure 1) positioned back-to-back and separated by a temporary wall. The EyeGaze setup does not incorporate sound, thus sound was provided as is within the room.

Face-to-Face
The participants were seated facing each other across a table. No technology was used to facilitate the discussion.

Skype
The Skype setup consisted of two connected laptops, each with a webcam build into the bezel of the screen. Both laptops were running an instance of the Skype application to transmit video to the other screen. The two laptops were placed in the same room, separated by a temporary wall. Audio was not provided via the Skype application, but as is within the room.

Test Setup

During the study, two rooms were used. One room was used for testing all three conditions, and an adjacent room for filling out a questionnaire. We constructed a specific EyeGaze setup, duplicated for each of the two test locations. These locations consisted of a 42" flatscreen LCD display with three Kinects placed stategically around the screen. In order to

maximise the captured area of the person sitting in front of the screen, a single Kinect was placed on each side of the TV pointing inwards, thereby allowing EyeGaze to capture each side of the person. The last Kinect was placed above the screen pointing downwards at a 20° angle.

Participants

30 participants were recruited by mailing students in the School of Information and Communication Technology and through Facebook. All participants except one were in their 20's and all spoke Danish fluently. Participants were randomly assigned a conversation partner according to the time they signed up. In total there were 30 participants, with eight being women.

Procedure

All test subjects participated in all three conditions with the same partner, and were afterwards asked to answer a questionnaire regarding their experiences. The order in which pairs of participants experienced the different conditions was randomised between tests. Before each test, the pair was given a list of four different subjects (nuclear power in Denmark, assisted suicide, a ban on fighting dogs, and mandatory usage of bicycle helmets), and asked to agree on which subject they would discuss, and who would argue for and who against the chosen suggestion. The pair of participants were then randomly placed in one of the three conditions, where they would discuss the subject for five minutes. The test leader would then ask them to pause the discussion and move to the next random condition, where they would continue the discussion. Five minutes later, they would move to the last condition. After the participants had discussed the chosen subject for 15 minutes, the discussion was interrupted by the test leader and the participants were asked to answer an online questionnaire, which was set up beforehand in an adjacent room. If the discussion came to an end before the fifteen minutes had gone by, the test leader gave a new subject for discussion.

Questionnaire

The questionnaire was mainly based on questions from Garau et al. [4].

Eye Contact

1. I had a good sense of eye contact with my conversation partner

Communication

2. I could easily tell when my partner was listening to me
3. I was able to take control of the conversation when I wanted to
4. It was easy for me to contribute to the conversation
5. The conversation seemed highly interactive
6. There were frequent and inappropriate interruptions
7. This felt like a natural conversation

Turn-Taking

8. I feel I interrupted my conversation partner often
9. I was often unsure of when my conversation partner was done talking

Involvement

10. I found it easy to keep track of the conversation
11. I felt completely absorbed in the conversation
12. I was easily distracted from the conversation

Co-Presence

13. I had a real sense of personal contact with my conversation partner
14. I was very aware of my conversation partner

Partner Evaluation

15. My partner was friendly
16. My partner did not take a personal interest in me
17. I trusted my partner
18. I enjoyed talking to my partner

Attentiveness

19. My conversation partner seemed attentive
20. My conversation partner listened to me

For each statement the users were asked to evaluate each condition on a nine-point Likert-scale with labels for each extreme ranging from ”Completely Disagree” to ”Completely Agree”. Furthermore, the users were asked whether they preferred EyeGaze or Skype on the particular condition by having a nine-point scale ancored at Skype at one end and EyeGaze at the other. Note that, negatively phrased questions have been normalised in the data analysis, such that a higher value is considered a more positive result.

FINDINGS

In this section we present the findings from our study questionnaire. We begin by presenting the findings from the comparison between Skype and EyeGaze. We then give a quick overview of the data collected in the questionnaire, which shows the general tendency found in the data. From there, we dig into the questionnaire data on each major questionnaire topic.

Comparison Between Skype and EyeGaze

The distinct difference between Skype and EyeGaze is that EyeGaze is facilitates eye contact between the two users. It is therefore not unreasonable to expect that users would prefer EyeGaze to Skype when discussing. Figure 2 presents the mean values of the participants' preference between Skype and EyeGaze for each question. In 18 out of 20 questions, the figure shows a preference for EyeGaze. Two exceptions to this trend are Question 7, which asks the participants if they felt the conversation was natural, and Question 12, which asks if the participants felt they were easily distracted from the conversation.

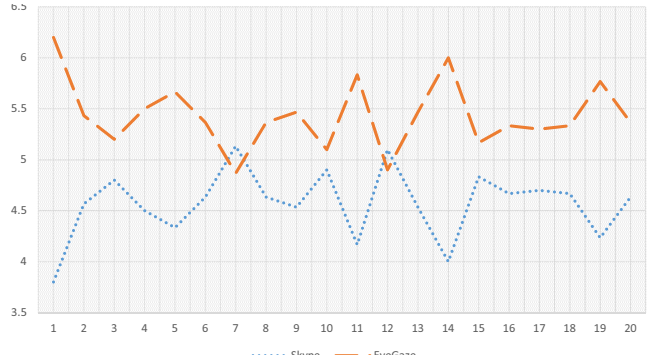


Figure 2: Participants' preference between Skype and EyeGaze per question

To test for significance, we averaged all questions into a single measure on the preference of Skype and EyeGaze. We used this measure to perform a One-Way ANOVA test, which showed a significantly higher rating for EyeGaze, $F(1, 19) = 33.17, p < 0.01$. In addition, we performed the One-Way ANOVA test on each individual question, to better understand in which areas EyeGaze was considered better.

In total, there were four questions which showed a significant difference between EyeGaze and Skype. Firstly, participants rated EyeGaze significantly higher than Skype in when asked the user if they had a good sense of eye contact (Question 1), $F(1, 29) = 6.64, p < 0.05$. This is an important result, as it shows that the participants noticed the presence of eye contact in EyeGaze, and the lack thereof in Skype. Secondly, when asked if the participants felt absorbed in the conversation (Question 11), participants rated EyeGaze significantly higher than Skype, $F(1, 29) = 10.76, p < 0.01$. This means that participants felt significantly more absorbed in the discussion when using EyeGaze than Skype.

EyeGaze was also rated significantly higher than Skype when asked if the participants were very aware of their conversation partner (Question 14), $F(1, 29) = 6.8, p < 0.01$. Finally, the last question with a significant difference between Skype and EyeGaze was Question 19, $F(1, 29) = 4.28, p < 0.05$, which asks the participant if their conversation partner seemed attentive. In this question, EyeGaze was also rated higher than Skype. This indicates that the presence of eye contact in EyeGaze helps the participants communicate attentiveness better.

For the rest of the questions there was a non-significant difference between EyeGaze and Skype. This also includes Questions 7 and 12, which rated Skype higher than EyeGaze.

Response Tendencies

We will now look at the findings from the participants' evaluation of each condition. A quick summary of the overall results can be seen in Figure 3 which shows the means of each condition for each question. We see from the means of the Skype and EyeGaze responses that there is a general tendency for EyeGaze to result in a slightly higher rating than Skype which alludes to the merits of EyeGaze. When comparing Skype and EyeGaze, the participants never rated Skype higher than EyeGaze. This points to better communication when eye contact is present, which is facilitated in EyeGaze. It is interesting to note that the mean values for Questions 15, 16, and 17 become very close to each other. These are questions on Partner Evaluation, and indicate an area where all three forms of communication perform very closely. The Face-to-Face condition consistently results in a higher rating on all questions compared to the EyeGaze and Skype condition, which is to be expected. EyeGaze was designed specifically to mimic the experience of eye contact in face to face communication. Thus, we classify two scenarios of positive results: *(i)* results which show a non-significant difference between Face-to-Face and EyeGaze, while showing a significant difference between Face-to-Face and Skype, *(ii)* and results which show a significant difference between Skype and EyeGaze.

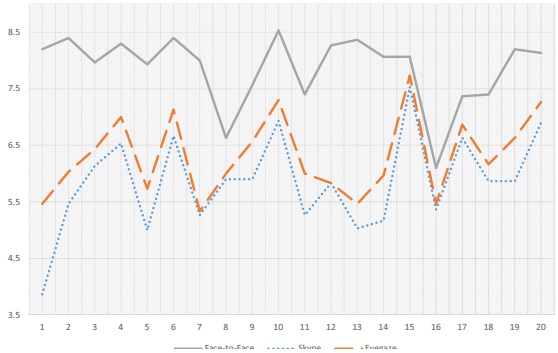


Figure 3: Summary of the findings for each condition over all questions

Eye Contact

The central argument for EyeGaze is that it facilitates the experience of eye contact. Figure 4 shows the mean responses from all 30 participants when asked if the participants felt they had a good sense of eye contact (Question 1). In this question, the participants on average rated EyeGaze to provide a better sense of eye contact than Skype, which was rated worst. The Face-to-Face condition was rated the highest.

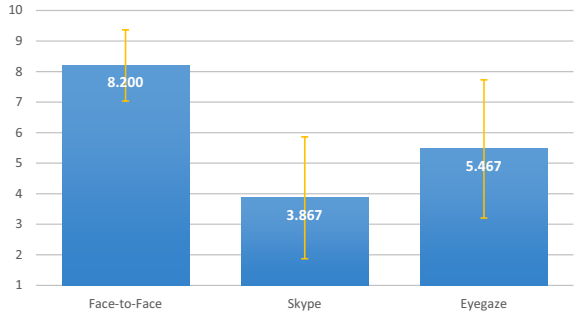


Figure 4: Response means for Question 1 showing that the mean rating of Skype was significantly lower than that of EyeGaze and Face-to-Face

To test for significance, we ran a One-Way ANOVA test, which showed a significant difference among the three conditions, $F(2, 58) = 42.84, p < 0.01$. A Tukey HSD post hoc test showed a significant difference between EyeGaze

and Skype ($p < 0.01$). This further strengthens the important result in the comparison between Skype and EyeGaze that the participants felt they achieved a better sense of eye contact using EyeGaze. It indicates that EyeGaze is capable of providing users with a sense of eye contact, which the participants largely agree that Skype is unable to provide. In addition, there was a significant difference between Face-to-Face and Skype ($p < 0.01$), and Face-to-Face and EyeGaze ($p < 0.01$). This shows that neither the Skype or EyeGaze condition performed as well as Face-to-Face condition.

Involvement

Eye contact has been shown to facilitate a better experience of involvement in the conversation. Figure 5 shows the mean values, when asked how absorbed they felt in the conversation (Question 11). A One-Way ANOVA test shows a significant difference among the three conditions, $F(2, 58) = 25.93, p < 0.01$. Firstly, the Tukey HSD post hoc test showed that EyeGaze was rated significantly higher than Skype ($p < 0.05$), which confirms the significant difference found in the comparison of Skype and EyeGaze. This points to better immersion between the participants when the experience of eye contact is facilitated. In addition, the Tukey HSD post hoc test showed a significant difference between Face-to-Face and Skype ($p < 0.01$) and Face-to-Face and EyeGaze ($p < 0.01$).

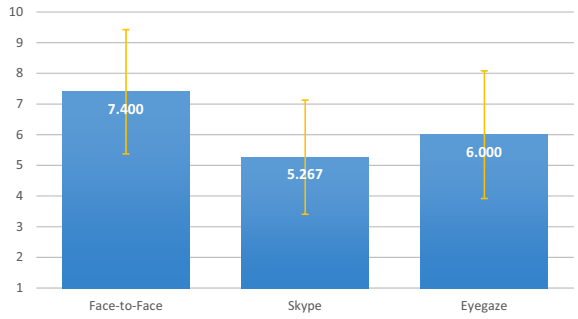


Figure 5: Response means for Question 11 in the Involvement measure

For the remaining two questions in the Involvement measure, the effect of eye contact is less clear. We performed the One-Way ANOVA test on Question 10, which asks the participant if they found it easy to keep track of the conversation, and Question 12, which asks if the participants were easily distracted from the conversation. The test showed a significant difference among the three conditions in Question 10, $F(2, 58) = 15.18, p < 0.01$, however the Tukey HSD post hoc test showed no significant difference between Skype and EyeGaze. There was, however, a significant difference between Face-to-Face and Skype ($p < 0.01$), and Face-to-Face and EyeGaze ($p < 0.01$). This indicates that neither Skype or EyeGaze allowed the participants to keep track of the conversation as well as the Face-to-Face condition.

Likewise for Question 12, the One-Way ANOVA test showed a significant difference amongst the three conditions, $F(2, 58) = 24.37, p < 0.01$. Again, the Tukey HSD post hoc test showed a non-significant difference between EyeGaze and Skype, while showing a significant difference between Face-to-Face and Skype ($p < 0.01$) and Face-to-Face and EyeGaze ($p < 0.01$). Similarly, this shows that neither the Skype or EyeGaze condition performed as well as the Face-to-Face condtion.

Partner Evaluation

How people are perceived is impacted by eye contact and gaze direction, as people sharing high amounts of mutual gaze seem to like each other more. Thus, allowing people to attain eye contact gives a venue for expressing liking of the other participant, and for observing signals sent by the other's gaze behaviour. Figure 6 shows the mean values when asked how friendly their conversation partner seemed (Question 15).

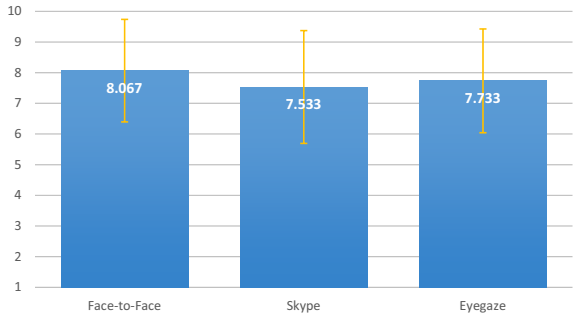


Figure 6: Response means of Question 15 in the Partner Evaluation measure

A One-Way ANOVA test showed a significant difference in the three conditions, $F(2, 58) = 6.66, p < 0.01$. The Tukey HSD post hoc test shows no significant difference between Face-to-Face and EyeGaze, while at the same time showing a significant difference between Face-to-Face and Skype ($p < 0.01$). This is consistent with aforementioned observations, and indicates that EyeGaze, which facilitates the experience of eye contact, allows the users to express liking and other signals better than Skype. There was no significant difference between Skype and EyeGaze.

This finding is mirrored when asked if the participants felt their conversation partner took no personal interest in them (Question 16). Again, the One-Way ANOVA test showed a significant difference in the three conditions, $F(2, 58) = 3.42, p < 0.05$. The Tukey HSD post hoc test also showed no significant difference between Face-to-Face and EyeGaze, while showing a significant difference between Face-to-Face and Skype ($p < 0.05$). There was no significant difference between Skype and EyeGaze. This further strengthens the argument that EyeGaze better allows for realistic partner evaluation.

For the remaining questions within the Partner Evaluation measure, the effects of eye contact do not show the same clear results. Question 17 asks if the participant trusted their partner. The One-Way ANOVA test showed a significant difference among the three conditions, $F(2, 58) = 8.26, p < 0.01$. Here, the Tukey HSD post hoc test showed no significant difference between Skype and EyeGaze. There was, however, a significant difference between Face-to-Face and EyeGaze

($p < 0.05$) and Face-to-Face and Skype ($p < 0.01$). This shows that neither Skype or EyeGaze performed as well as the Face-to-Face condition

Question 18 asks if the participant enjoyed speaking to their conversation partner. Again, the One-Way ANOVA test showed a significant difference among the three conditions, $F(2, 58) = 17.85, p < 0.01$. Like the previous question, the Tukey HSD post hoc test show a non-significant difference between Skype and EyeGaze. In addition, there was a significant difference between Face-to-Face and Skype ($p < 0.01$) and Face-to-Face and EyeGaze ($p < 0.01$). Again, this shows that neither the Skype or EyeGaze condition performed as well as the Face-to-Face condition.

Turn-Taking

With a sense of eye contact comes several benefits such as the ability to moderate the conversation and indicate when it is your turn to speak. The mean values for the three conditions when asked how much the participant felt they interrupted their conversation partner (Question 8) is shown in Figure 7 The figure shows that EyeGaze was rated below Skype, which indicates that the participants were able to better mediate the conversation when using EyeGaze. Note that this question is phrased negatively, thus lower values are considered better.

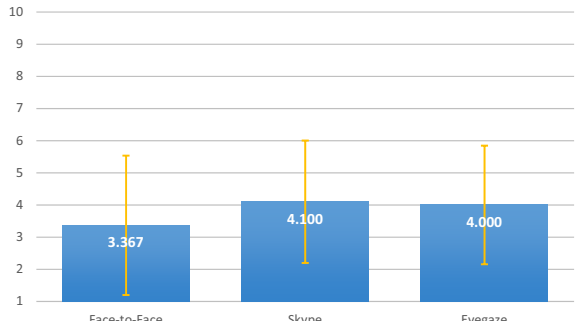


Figure 7: Response means for Question 8 in the Turn-Taking measure. Mean ratings of EyeGaze was not significantly higher than Face-to-Face

A One-Way ANOVA test performed on Question 8 shows a significant difference between the three conditions, $F(2, 58) = 3.76, p < 0.05$. The Tukey HSD post hoc test shows a non-significant difference between Face-to-Face and EyeGaze, while showing a significant difference between Face-to-Face and Skype ($p < 0.05$). This points to the participants having a better flow experience when in the EyeGaze condition compared to when the discussion was held in the Skype condition. There is no significant difference between Skype and EyeGaze.

The effects of eye contact are not as clear on the second question in Turn-Taking. The second question in Turn-Taking asks the participants if they were often unsure of when their conversation partner was finished talking (Question 9). The One-Way ANOVA test showed a significant difference among the three conditions, $F(2, 58) = 9.29, p < 0.01$. The Tukey HSD post hoc test showed no significant difference between Skype and EyeGaze. There was a significant difference between Face-to-Face and Skype ($p < 0.01$), and Face-to-Face and EyeGaze ($p < 0.05$). This indicates that neither the Skype or EyeGaze condition performed as well as the Face-to-Face condition.

Attentiveness

Higher levels of face gaze give the impression of being more attentive than lower levels. Thus, we would expect a difference between Skype and EyeGaze in EyeGaze's favour, as it is the only one of the two that permits mutual face gaze. Question 19, asks participants if they thought their partner seemed attentive, and Question 20 if their partner seemed to listen to them. We have summarised both into a single combined measure as both questions show the same results. Figure 8 shows the means for the combined measure of Attentiveness.

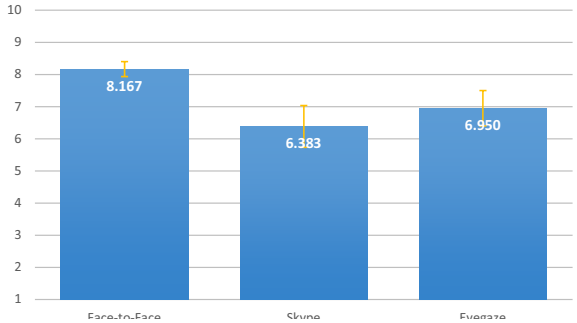


Figure 8: Response means for the measure of attentiveness showing a slight, but non-significant bias towards EyeGaze compared to Skype

A One-Way ANOVA analysis shows a significant difference between the three conditions, $F(2, 58) = 24.62, p < 0.01$. While there is a slight bias towards EyeGaze, the Tukey HSD post hoc test shows it is not significant. There is, however, a significant difference between Face-to-Face and Skype ($p < 0.01$) and Face-to-Face and EyeGaze ($p < 0.01$). This indicates that neither the Skype condition, or the EyeGaze condition were able to convey the experience of attentiveness as well as the Face-to-Face condition.

Since there is a significant difference in the comparison between Skype and EyeGaze in Question 19, it is reasonable to expect a significant difference in the rating of EyeGaze and Skype. A One-Way ANVOA analysis on Question 19 reveals a significant difference among the three conditions, $F(2, 58) = 21.75, p < 0.01$, however, the Tukey HSD post hoc test reveals no significant difference between Skype and EyeGaze, which is surprising. There is a significant difference between Face-to-Face and Skype ($p < 0.01$) as well as Face-to-Face and EyeGaze ($p < 0.01$), which as with the combined measure, indicates that neither Skype or EyeGaze performed as well as the Face-to-Face condition.

Communication

Questions in the Communication measure reflect on all aspects of communication, from the flow of the conversation, to how natural the conversation felt. The findings in this measure all show similar, unclear results. For this reason, we collate and present them as the single measure Communication. Figure 9 shows the mean ratings for each condition on

the Communication measure. The figure indicates a general, but small bias toward EyeGaze compared to Skype. Performing a One-Way ANOVA test showed a significant difference within the three conditions, $F(2, 58) = 42.7, p < 0.01$ However, the results of a Tukey HSD post hoc test show that there is no significant difference between Skype and EyeGaze, and a significant difference between both Face-to-Face and Skype ($p < 0.01$), and Face-to-Face and EyeGaze ($p < 0.01$). This indicates that the Skype and EyeGaze condition did not provide natural setting and ability to regulate conversation flow as well as Face-to-Face in the Communication measure.

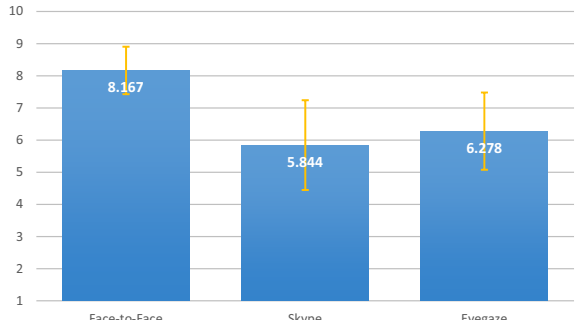


Figure 9: Response means for the Communication measure

Co-Presence

The two questions which make up the Co-Presence measure pertain to the feeling of having achieved a sense of personal contact with conversation partner. The amount of eye contact between conversation partners has an influence on the impression they get of each other. Thus, it would be expected that eye contact would attribute to a sense of personal awareness of a conversational partner. As with Attentiveness and Communication, we have summarised Question 13, which asks the participant if they felt a real sense of personal contact with their conversation partner, and Question 14, which asks the participant if they were very aware of conversation partner, into the Co-Presence measure. Figure 10 shows the mean ratings for each condition on the Co-Presence measure.

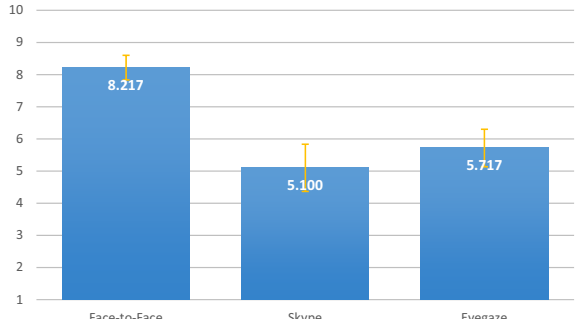


Figure 10: Response means for the Co-Presence measure

A One-Way ANOVA analysis showed a significant difference among the three conditions, $F(2, 58) = 48.25, p < 0.01$. The Tukey HSD post hoc test shows, however, that like Attentiveness and Communication, there is no significant difference between Skype and EyeGaze, while both Face-to-Face and Skype ($p < 0.01$) and Face-to-Face and EyeGaze ($p < 0.01$) are rated significantly lower than Face-to-Face. This indicates that adding eye contact does not have a significant effect on a persons awareness of their conversational partner's presence.

DISCUSSION

In this section we discuss the significance of our findings. Lastly we discuss the limitations of EyeGaze.

Evaluation of Findings

Our study showed a general tendency to rate EyeGaze slightly better than Skype, however, when looking at the overall findings we found a surprisingly small difference between EyeGaze and Skype, compared to what we had expected.

Comparison Between Skype and EyeGaze

Looking specifically at the direct comparison between the Skype and EyeGaze conditions we see that EyeGaze is indeed able to provide a more satisfactory experience. As our findings showed, there is a significant difference between EyeGaze and Skype when combining all questions into a single measure. When we look at the questions separately, we see that EyeGaze is rated higher than Skype in 18 out of 20 questions, significantly so in four questions. Thus, users appear to prefer EyeGaze over Skype on most issues - both those concerning the experience of the conversation and the partner. The two expections to this are when asked if it felt like a natural conversation (Question 7) and if they were easily distracted from the conversation (Question 12). Both questions could be influenced by the fact that EyeGaze was a novel experience for users. Some users were observed to look at the Kinect cameras relatively often, which might both distract and make the conversation feel less natural for both participants. Furthermore, several participants had used Skype for communicating via video before, making that condition more natural for them. Overall, though, EyeGaze appears to be able to provide a more natural conversation than that of Skype.

Eye Contact

The central objective in the design and implementation of the EyeGaze system was to facilitate a better understanding of eye contact in a video mediated setting. When asked directly on the level of eye contact in the three conditions comparatively, our findings showed that the participants had a significantly better sense of eye contact using EyeGaze than using the Skype condition. This can be attributed to the corrected view point in the EyeGaze system which corrects the downwards pointing gaze of the Skype condition. By mapping the texture onto our real-time model and using a personalised view point, the findings shows that we can indeed obtain a form of eye contact and that this eye contact is significantly better than that provided by the wide spread Skype application.

It is worth noting that even though the participants were directly asked about eye contact and keeping in mind the fact

that it is not technically possible to achieve eye contact between two individuals through Skype, the mean response was still relatively high (3.867 ± 2.030) for the Skype condition. This might be caused by participants looking directly into the camera, with the purpose of giving the conversation partner a sense of eye contact - though they themselves would not be able to see the other person simultaneously. An alternative explanation comes from Grayson & Monk [5] who showed that people are able to relatively quickly train themselves to understand where another person is looking and when that person was trying to establish face gaze when looking at a video feed off screen.

Involvement

The significance of the question on how absorbed the participant felt in the conversation (Question 11) shows that the possibility of establishing two-way eye contact has a profound impact on the involvement in the conversation. It indicates that the possibility of eye contact makes it easier to immerse oneself in the conversation. This is supported by Scherer [17], who found that conversation partners found each other more interesting when there were higher amounts of eye contact. This points to eye contact having a impact on how people are able to cope with distractions, which Questions 10 and 12 indicated that there were significantly more of in Skype and EyeGaze. The high amount of interruptions in both Skype and EyeGaze could point towards some conversational moderating abilities being lost in both these conditions compared to the Face-to-Face condition.

Partner Evaluation

Questions pertaining to Partner Evaluation (Question 15 to 18) deal with the test participants' perception of their conversation partner. The first two questions showed no significant difference between Face-to-Face and EyeGaze, while showing a significant difference between Face-to-Face and Skype. This indicates that EyeGaze gives a more realistic impression of how friendly etc. the conversation partner is than Skype. As Skype left users feeling significantly less involved than in both Face-to-Face and using EyeGaze, the Partner-Evaluation measure might, similar to involvement be influenced by the sense of interest facilitated by the experience of EyeGaze.

The last two questions in this measure showed significant differences between both Face-to-Face and EyeGaze, and Face-to-Face and Skype. This is especially understandable in the last Question 18, which asks about enjoyment. Both the individual test setups and the context can have impacted the perceived enjoyment of engaging in the discussion. As we mentioned in our findings, the first three questions in this measure all showed very similar results for all three conditions. This could indicate that this is an area in human communication which all three conditions are capable of relaying relatively well.

Turn-Taking

There was a significant difference between Face-to-Face and Skype, but not Face-to-Face and EyeGaze, when asking users if they interrupted their conversation partner often (Question 8). This indicates that adding eye contact to a system faciliates a more pleasant turn taking experience, like that which is acheived in normal face-to-face communication. The significant difference between Face-to-Face and Skype shows that simply being able to see the user is not enough to facilitate this naturally and that more is required, such as eye contact. Possible reasons for the lack of a significant difference between EyeGaze and Skype could potentially be the differences in proxemics and interpersonal communication setting. When asking users if they were often unsure of when their partner was done talking (Question 9), we found a significant difference between both Face-to-Face and Skype, and Face-to-Face and EyeGaze. Thus, it seems users were unsure of when it was their turn to speak in the two video-mediated conversations compared to when talking face-to-face, but it did not result in a significant difference for EyeGaze in how much they felt they interrupted their partner compared to Face-to-Face. People might be more hesitant to begin speaking when unsure of if the other is still talking, explaining how these findings suggest that users of EyeGaze were more unsure of when their partner was done than in Face-to-Face, but did not interrupt significantly more than in Face-to-Face.

Attentiveness

The Attentiveness measure is comprised of two questions: one if their conversation partner seemed attentive (Question 19), and one if they felt their partner listened (Question 20). Again, there is a bias toward EyeGaze opposed to Skype, but with no significant difference. Research has shown that people are found more attentive when they maintain eye contact while listening. This is not confirmed by our findings, where there is a significant difference between Face-to-Face and EyeGaze, which both facilitate a sense of eye contact. There is, furthermore, a significant difference between Face-to-Face and Skype in both questions, as well as the combined measure. However, the comparative part of Question 19 revealed a significant difference between EyeGaze and Skype, suggesting that when asking the participants to comparing the two directly, participants report a greater difference than when rating them separately. It bears mentioning that the Involvement category might be related to the Attentiveness category, in the sense that users who feel more distracted in a conversation could be expected to also appear less attentive. Thus, this might have impacted the findings in regards to Question 19 and Question 20.

Communication and Co-Presence

The findings become even less positive for EyeGaze when looking at the following two measures, Communication and Co-Presence. They both show a significant difference between Face-to-Face and EyeGaze, and Face-to-Face and Skype, while showing no significant difference between Skype and EyeGaze. This finding is interesting and points to a general challenge when trying to evaluate the impact of eye contact in this setting. It can be argued that the use of eye contact is simply a single part of a much larger toolbox that we use in interpersonal communication. This can make it

difficult to isolate eye contact as the contributing factor. In addition, this makes measuring the benefits of eye contact challenging, as the benefits can be masked. However, the average means of EyeGaze are better than Skype across both measures, mirroring the general tendency toward Skype scoring better than EyeGaze.

Proxemics and Interpersonal Communication

One issue with EyeGaze which could have impacted negatively on the participants' rating of their experience is the proxemics involved in the conversation. The EyeGaze system was setup to facilitate a 1:1 correspondence between the actual distance between the participants and the rendered distance. Due in part to the seating arrangement of the setup, and due to the Xbox 360 Kinect's minimum depth range of 80 cm, the perceived distance between the two participants was between 2 - 2.5 meters. A close phase social distance between participants as defined by Hall [6] is 1.2 - 2.1 metres, meaning that the perceived distance between subjects in the EyeGaze setup risked placing the other at a distance usually reserved for speeches and lectures. While the far phase social distance is describe by Hall as between 2.1 and 3.6 metres, this distance is typically described as the distance you move when someone tells you to step away so they can look at you. In the context of video mediated conversation, it therefore makes more sense to compare with the close phase social distance. This might inadvertently change the participants' expectations and reactions to each other and the conversation and is in stark contrast with the more intimate distance of the Skype condition.

Limitations

Since Face-to-Face has achieved high ratings in our study, we do not interpret the results to indicate that eye contact and face gaze are unimportant in a video mediated setting. Instead, the reasons behind the relatively similar ratings between Skype and EyeGaze are likely to be found in the EyeGaze system itself, which, due to its prototype nature, is still relatively crude. The frame rate of EyeGaze is still quite low compared to normal video frame rates. Using all three Kinect cameras and a network connection for passing skeleton data, the EyeGaze system was running at approximately 8 frames per second which reduces the feeling of presence. Since model update and render frame rates are linked, the poor frame rate also means that the underlying 3D model is updated slower, which can provide unsatisfactory intermediate image if the user moves frequently and quickly.

When using commodity cameras and 3D-sensors, there is a trade-off in quality for convenience. Amongst the problems caused by the commodity hardware was the quality of the skeletal tracking in a seated position. Some users experienced unexpected jitter in the rendered perspective of the video, as the internal skeletal tracking miscalculated the position of the user's head or failed to recognise the skeleton for a moment. This type of jitter breaks the illusion that the screen is a window. To solve this problem it might be advisable to employ a more stable approach to head tracking, such as recognition of facial features in the RGB image.

Some test participants whose conversation partner wore glasses reported having issues with the EyeGaze program. The eyes of the subject wearing glasses did not reliably sit in the correct place of the subject's head, which naturally made it difficult for the viewer to achieve a realistic sense of eye contact.

Focus was given to the implementation of eye contact in EyeGaze and hence sound was ignored in the implementation. As a conversation was an important part of our study setup, both participants were seated in the same room, separated by a thin, movable wall. Thus, the sound was experienced in the same way in all three conditions: as-is in the room and coming from directly across the table from the user. This might potentially reflect the worst on the EyeGaze condition, as the higher processing requirements could lead to synchronisation issues between sound and video. In fact, many participants commented after the study that they answered some questions based on the experience of the sound, and not the image and video representation. Attempts at removing these issues are not expected to be simple.

No efforts were made to ensure a realistic demographic spread of test participants. Due to the nature of the recruitment effort, most test subjects were in fields related to computer science. This, along with their age, might have made a difference for their tolerance of the faults in the EyeGaze program, and might also affect the possibility of being an experienced user of Skype.

CONCLUSION

In this paper we have presented a within-subject study using 30 participants, in three conditions, to evaluate the importance of eye contact in a video mediated communication setting. To evaluate this we have developed EyeGaze, an eye contact-enabled video communication system, which captures the user in 3D and allows a virtual viewpoint to be positioned correctly relative to an observers personal viewpoint. We have presented literature on the importance of eye contact in human communication and the functions that eye contact facilitates. We then presented the results from our study which was based on questionnaire responses from all 30 participants, where they were asked to rate each question in relation to each condition, and judge whether they found Skype or EyeGaze more pleasing in relation to the question. Our study showed three trends: *1*) EyeGaze was rated significantly better than Skype; *2*) EyeGaze was not rated significantly worse than Face-to-Face, while Skype was rated significantly worse than Face-to-Face, and finally; *3*) both Skype and EyeGaze were rated significantly worse than Face-to-Face.

Our findings show that users rated EyeGaze significantly higher than Skype, when asked if they felt they had a good sense of eye contact. This is an important result, as facilitating an experience of eye contact is the primary goal in the design of EyeGaze. In addition our study showed a general tendency towards a better experience in EyeGaze compared to Skype, however, this tendency was not always significant. Both EyeGaze and Skype performed significantly worse than Face-to-Face, which was expected.

FUTURE WORK

Potential future work would be to generalise the study presented to see if the results could be confirmed. It would be interesting to see whether the general tendency to rate EyeGaze higher than Skype would achieve statistical significance when performing a similar study with more participants.

REFERENCES

1. Bee, N., Wagner, J., André, E., Vogt, T., Charles, F., Pizzi, D., and Cavazza, M. Discovering eye gaze behavior during human-agent conversation in an interactive storytelling application. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, ACM (2010), 9.
2. Duncan, S., and Niederehe, G. On signalling that it's your turn to speak. *Journal of Experimental Social Psychology 10*, 3 (1974), 234–247.
3. Fry, R., and Smith, G. F. The effects of feedback and eye contact on performance of a digit-coding task. *The Journal of Social Psychology 96*, 1 (1975), 145–146.
4. Garau, M., Slater, M., Bee, S., and Sasse, M. A. The impact of eye gaze on communication using humanoid avatars. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM (2001), 309–316.
5. Grayson, D. M., and Monk, A. F. Are you looking at me? eye contact and desktop video conferencing. *ACM Trans. Comput.-Hum. Interact. 10*, 3 (Sept. 2003), 221–243.
6. Hall, E. T. *The Hidden Dimension*. Anchor, Oct. 1990.
7. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., and Fitzgibbon, A. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST '11, ACM (New York, NY, USA, 2011), 559–568.
8. Jensen, A. K., Nielsen, T. S., and Smedegård, J. H. EyeGaze: Eye Contact and Gaze Awareness in Video. 2013.
9. Kendon, A. Some functions of gaze-direction in social interaction. *Acta Psychologica 26*, 0 (1967), 22 – 63.
10. Kipp, M., and Gebhard, P. Igaze: Studying reactive gaze behavior in semi-immersive human-avatar interactions. In *Intelligent Virtual Agents*, Springer (2008), 191–199.
11. Kleinke, C. L. Gaze and Eye Contact: A Research Review. *Psychological Bulletin 100*, 1 (1986), 78–100.
12. Kleinke, C. L., Bustos, A. A., Meeker, F. B., and Staneski, R. A. Effects of self-attributed and other-attributed gaze on interpersonal evaluations between males and females. *Journal of Experimental Social Psychology 9*, 2 (1973), 154 – 163.
13. Kleinke, C. L., Meeker, F. B., and Fong, C. L. Effects of gaze, touch, and use of name on evaluation of engaged couples. *Journal of Research in Personality 7*, 4 (1974), 368 – 373.
14. Kraut, R. E., and Poe, D. B. Behavioral roots of person perception: The deception judgements of customs inspectors and layment. *Journal of Personality and Social Psychology 39*, 5 (1980), 784–798.
15. Matthias, R., and André, E. From chatterbots to natural interaction–face to face communication with embodied conversational agents. *IEICE transactions on information and systems 88*, 11 (2005), 2445–2452.
16. Mehrabian, A. Inference of attitudes from the posture, orientation and distance of a communicator. *Journal of Consulting and Clinical Psychology 32*, 3 (1968), 296–308.
17. Scherer, S. E. Influence of proximity and eye contact on impression formation. *Perceptual and Motor Skills 38*, 2 (April 1974).
18. Thayer, S., and Schiff, W. Observer judgement of social interaction: Eye contact and relationship inferences. *Journal of Personality and Social Psychology 30*, 1 (1974), 110–114.

Bibliography

[1] Garau, M., Slater, M., Bee, S., and Sasse, M. A. The impact of eye gaze on communication using humanoid avatars. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM (2001), 309–316.

[2] Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., and Fitzgibbon, A. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST '11, ACM (New York, NY, USA, 2011), 559–568.

[3] Maimone, A., and Fuchs, H. Encumbrance-free telepresence system with real-time 3d capture and display using commodity depth cameras. In *Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality*, ISMAR '11, IEEE Computer Society (Washington, DC, USA, 2011), 137–146.

[4] Paay, J., Kjeldskov, J., and O'Hara, K. Bisi: a blended interaction space. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, ACM (New York, NY, USA, 2011), 185–200.

[5] Usoh, M., Catena, E., Arman, S., and Slater, M. Using presence questionnaires in reality. *Presence: Teleoperators & Virtual Environments 9*, 5 (2000), 497–503.